UMUTeam at SemEval-2024 Task 4: Multimodal Identification of Persuasive Techniques in Memes through Large Language Models

Ronghao Pan[1], **José Antonio García-Díaz**[1], **Rafael Valencia-García**[1]
[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan, joseantonio.garcia8, valencia}@um.es

Abstract

In this manuscript we describe the UMUTeam's participation in SemEval-2024 Task 4, a shared task to identify different persuasion techniques in memes. The task is divided into three subtasks. One is a multimodal subtask of identifying whether a meme contains persuasion or not. The others are hierarchical multi-label classifications that consider textual content alone or a multimodal setting of text and visual content. This is a multilingual task, and we participated in all three subtasks but we focus only on the English dataset. Our approach is based on a fine-tuning approach with the pre-trained RoBERTa-large model. In addition, for multimodal cases with both textual and visual content, we used the LMM called LlaVa to extract image descriptions and combine them with the meme text. Our system performed well in three subtasks, achieving the tenth best result with an Hierarchical F1 of 64.774%, the fourth best in Subtask 2a with an Hierarchical F1 of 69.003%, and the eighth best in Subtask 2b with a Macro F1 of 78.660%.

1 Introduction

The rise of social media has facilitated the rapid spread of information. However, its unconstrained nature has also led to the spread of information whose accuracy is difficult to verify. As a result, misinformation and disinformation have become serious problems in everyday life. For example, during the COVID-19 pandemic, social media enabled healthcare professionals to quickly communicate professional information to the public; however, studies also revealed the spread of inaccurate health-related information (Ferrara et al., 2020).

A special case of spreading misinformation is the use of memes. Memes consist of images overlaid with text created by Internet users and have become one of the primary forms of content in online disinformation campaigns. Designed specifically to actively spread inaccurate information, "disinformation memes" are particularly effective on social media platforms, where they can quickly reach large audiences (Qu et al., 2022). Using various rhetorical and psychological techniques such as causal oversimplification, name-calling, and smear tactics, memes play a pivotal role in influencing users' perceptions and beliefs.

To address this phenomenon, the Multilingual Detection of Persuasion Techniques in Memes shared task has been organized at SemEval-2024 (Dimitrov et al., 2024). The goal of this task is to develop models for detecting persuasion techniques in the textual content of a meme, as well as in a multimodal setting where both textual and visual content are analyzed together. The task is divided into three main subtasks:

- **Subtask 1**. This is a unimodal hierarchical multi-label classification. The goal is to identify which of the 20 persuasion techniques are present using only textual features.
- **Subtask 2a**. This is a multimodal hierarchical multi-label classification. The goal is to identify which of the 22 persuasion techniques are present using textual and visual multimodal features.
- **Subtask 2b**. This is a multimodal binary persuasion identification task, where the goal is to determine whether a meme contains a persuasion technique or not.

To solve the English challenge, we propose an approach based on fine-tuning Transformer models for binary and hierarchical multi-label classification problems of persuasion techniques using textual and visual content. During training for subtasks 2a and 2b, we used a Large Multimodal Model (LMM) called LlaVa (Liu et al., 2023) to extract textual and visual features from the memes. We then refined the monolingual model, as RoBERTa-large (Liu et al., 2019), to identify persuasion techniques and their type.

In multimodal classification problems, our experiments showed that including the textual description of the meme obtained by an LMM improves the overall performance. In our experiments, this strategy achieved better results and required fewer resources than merging the image and text embeddings into the same vector space.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 655–666
June 20-21, 2024 ©2024 Association for Computational Linguistics

The rest of this paper is organized as follows. Section 2 provides a summary of important details about the task setup. Section 3 provides an overview of our system for two subtasks. Section 4 presents the specific details of our systems. Section 5 discusses the results of the experiments, and finally the conclusions are presented in section 7.

2 Background

Recently, there has been a significant increase in the use of memes on social media as a means of spreading misinformation. Memes consist of a combination of text and images that together have a meaning that is very difficult to automatically verify. In addition, the image and text of the meme in isolation may convey a benign meaning, but their combination may be derogatory, or vice versa. Fake news and hate speech purveyors use memes as a tool to spread misinformation and hateful content. They may spread hate to create unrest among the people, and such hateful content may target communities or individuals based on religion, ethnicity, race, national origin, affiliation, sexual orientation, gender, sex, disability, and disease (Hamza et al., 2023).

Many studies have focused on identifying memes that contain negative content or misinformation. For example, the authors of (Hamza et al., 2023) published a dataset of religiously hateful memes and evaluated it fine-tuning VisualBERT, which was pre-trained on the Conceptual Caption (CC) dataset for the top-down classification task. Visual features were extracted using ResNeXT-152 Aggregated Residual Transformations based Masked Regions with Convolutional Neural Networks (R-CNN) and BERT without textual encoding for the early fusion model. Regarding multimodal approaches, there have been tasks previously organized in the same area of interest. MAMI (Multimedia Automatic Misogyny Identification) at SemEval-2022 (Fersini et al., 2022), which explored the detection of misogynistic memes on the web using available text and images; and DravidianLangTech at EACL-2021 (Suryawanshi and Chakravarthi, 2021), which explored the detection of offensive language and classification of troll memes.

The novelty of this shared task is the focus on disinformation propaganda through memes. Propaganda uses psychological and rhetorical techniques to achieve its goal. These techniques include the use of logical fallacies and appeals to the audience's emotions. Logical fallacies are often difficult to detect because the argument seems correct and objective at first glance. However, careful analysis reveals that the conclusion cannot be deduced from the premise without the misuse of logical rules. Therefore, memes are a perfect medium for spreading disinformation because they consist of an image superimposed on text, and the image can be deceptive, reinforcing or complementing one or more persuasive techniques in the text or image. Thus, the goal of this task is to identify the existence and type of persuasion techniques through memes with different subtasks. The persuasion techniques can be viewed on the official task page.[1]. It is worth noting that a similar propaganda technique was used in Dipromats 2023 (IberLEF) (Moral et al., 2023).

The dataset used for this task is the one provided by the organizers. It consists of a set of texts and images labeled with their corresponding persuasion techniques and a binary annotation for Subtask 2b. The data set provided by the organizers is divided into train, dev, and validation. Note that we do not actually need two datasets for validation (dev and validation), since the dev set was used for the development phase. Therefore, we have combined the train and dev sets into a single training set. The training dataset contains 8000 examples for subtasks 1 and 2a and 1499 examples for subtask 2b. Figure 1 shows the distribution of the training set for subtasks 1a and 2a and Figure 2 for subtask 2b.

3 System overview

Figure 3 shows the architecture of our system for the three subtasks. We can see that for Subtask 1, only the text of the memes is used, which is a multi-label classification problem of different persuasion techniques. To address subtask 1, we have fine-tuned RoBERTa-large (Liu et al., 2019). For Subtasks 2a and 2b, we have used a similar approach as in Subtask 1, but including textual and visual features. We rely on LlaVa (Liu et al., 2023) to extract the image description and then concatenate this information with the textual content of the memes, as shown in Figure 4. LlaVa is an end-to-end multimodal Large Language Model (LLM) that incorporates a vision encoder for general pur-

[1] https://propaganda.math.unipd.it/semeval2024task4/

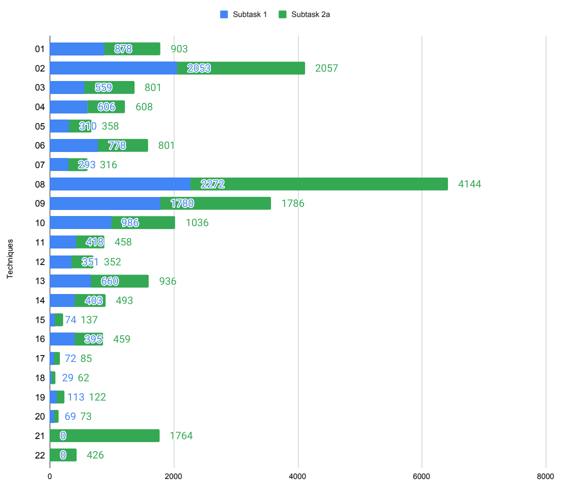

Figure 1: Distribution of training for Subtasks 1 and 2a. The techniques are: (01) Black-and-white Fallacy / Dictatorship; (02) Loaded Language; (03) Glittering generalities (Virtue); (04) Thought-terminating cliché; (05) Whataboutism, (06) Slogans, (07) Causal Oversimplification; (08) Smears; (09) Name calling/Labeling; (10) Appeal to authority; (11) Exaggeration/Minimisation; (12) Repetition; (13) Flag-waving; (14) Appeal to fear/prejudice; (15) Reductio ad hitlerum; (16) Doubt; (17) Misrepresentation of Someone's Position (Straw Man); (18) Obfuscation, Intentional vagueness, Confusion; (19) Bandwagon; (2) Presenting Irrelevant Data (Red Herring); (21) Transfer; (22) Appeal to (Strong) Emotions.

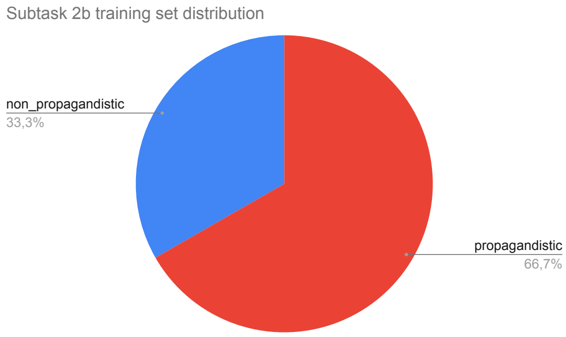

Figure 2: Distribution of training for Subtask 2b.

pose visual and language understanding. LlaVa has demonstrated impressive multimodal conversational capabilities, sometimes exhibiting behavior similar to the multimodal GPT-4 on unseen images/instructions, and achieving a relative score of 85.1% compared to GPT-4 on a synthetic multimodal instruction-following dataset (Liu et al., 2023). It is worth noting that the output model as a binary classification problem and in subtask 2b as a multi-class hierarchical classification problem like subtask 2a.

4 Experimental setup

In this work, we used only the dataset provided by the organizers and we did not rely on external data except for the use of LLM and LMM models that were pre-trained with general purpose data.

Before fine-tuning, we performed a preprocessing step to remove line breaks, hashtags, symbols, references, and hyperlinks. Next, for all the subtasks, we performed the fine-tuning process using an epoch-based evaluation strategy with the Huggingface Trainer library[2]. This involves training the pre-trained model for a certain number of epochs and performing an evaluation with the evaluation set after each epoch. Once all epochs have been completed, the model with the best macro F1 score in the evaluation set is selected. In this way, overfitting or underfitting resulting in low variance and high bias can be avoided.

We used the same hyperparameters for fine-tuning in all the subtasks: (1) a batch size of 8 for both training and validation, (2) 10 epochs, (3) a learning rate of 2e-5, (4) and a weight decay of 0.01. During training, we used macro-F1 as a reference. For the evaluation of subtasks 1 and 2a, the organizers used hierarchical-F1 as the primary evaluation metric, and for subtask 2b, macro-F1. It should be noted that in order to ensure the reproducibility of the experiment, we modified the LlaVa generation configuration by setting the value of `do_sample` to False.

Hierarchical precision, recall and F1 (H-P, H-R, and H-F1) are metrics used in hierarchical classification problems where classes are organized in a hierarchical structure (Kiritchenko et al., 2006). H-F1 considers both precision and recall of the prediction for each class in the hierarchy, taking into account the relationship between parent and child classes in the hierarchy.

The binary task (subtask 2b) is evaluated using the macro F1 score, which is an evaluation metric used in classification problems to measure the precision and recall of a model in predicting multiple classes. It assigns equal weight to each class, meaning that all classes have the same impact on the final metric, regardless of their size or distribution in the data.

[2] https://huggingface.co/docs/transformers/main_classes/trainer

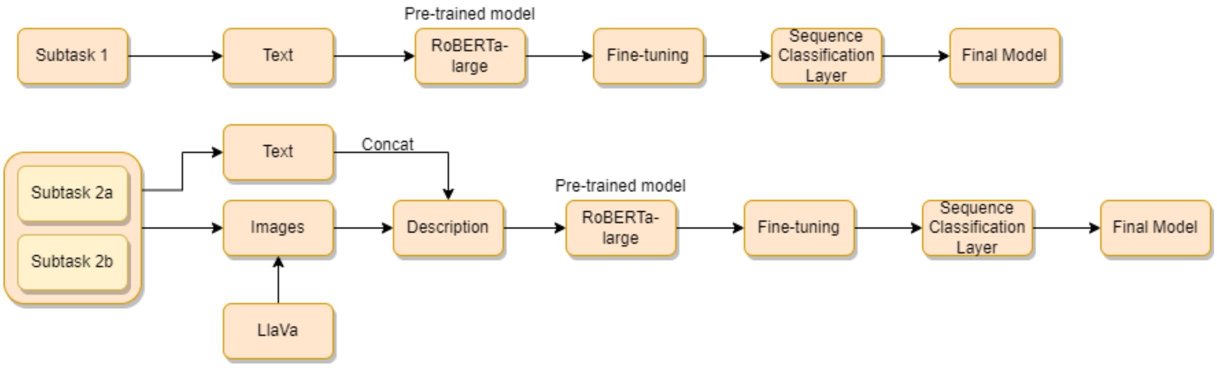

Figure 3: System architecture approach

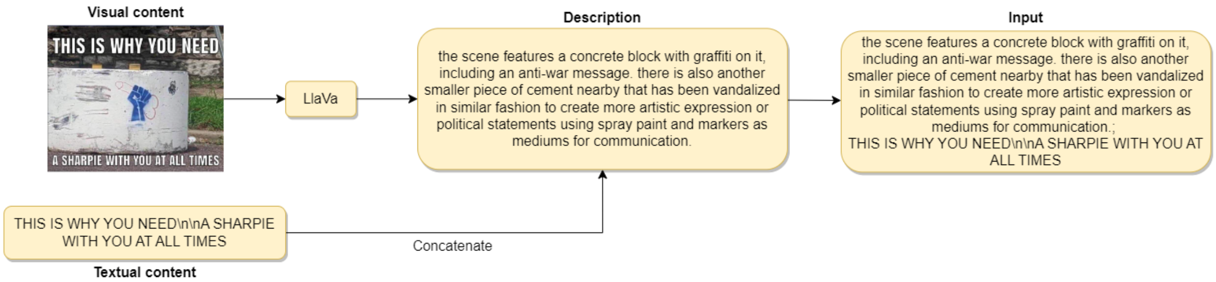

Figure 4: Example of multimodal input for Subtasks 2a and 2b.

5 Results

Table 1 shows the official ranking table for Subtask 1. We can see that we have ranked tenth position with a H-F1 score of 0.64774 and a H-P of 70.817%. We outperformed the baseline by almost 50% in terms of H-F1 and are 10.473% away from the best result, which is 75.247%, achieved by the '914isthebest' team.

Table 1: Official results for the Subtask 1.

Team	Rank	H-F1	H-P	H-R
914isthebest	1	75.247	68.419	83.590
BCAmirs	2	69.857	66.786	73.223
Otterly Obsessed With Semantics	3	69.738	64.801	75.490
TUMnlp	4	67.384	63.781	71.419
GreyBox	5	66.998	65.248	68.844
		...		
UMUTeam	10	**64.774**	**70.817**	**59.681**
		...		
Baseline	-	36.865	47.711	30.036

For Subtask 2a, we achieved a H-F1 of 69.003% and a H-P of 76.763%, which is the fourth-best result according to the official ranking table (see Table 2). With our approach, we outperformed the baseline by 24.297% and are only 5.589% away from first place, which achieved an H-F1 of 74.592%.

Table 2: Official results for the Subtask 2a.

Team	Rank	H-F1	H-P	H-R
Hierarchy Everywhere	1	74.592	86.682	65.461
NLPNCHU	2	70.677	78.164	64.498
BCAmirs	3	70.497	78.374	64.059
UMUTeam	4	**69.003**	**76.763**	**62.669**
		...		
Baseline	-	44.706	68.778	33.116

For subtask 2b, which is a binary classification problem to identify the presence of persuasion techniques in memes, we obtained a macro-F1 score of 78.660%, which puts us in eighth place according to the official ranking table (see Table 3). Furthermore, we can see that our system has improved by up to 53.66% compared to the baseline and is only 2.37% behind the first place (LMEME with a M-F1 of 81.030%).

Based on the results obtained, it's clear that combining image descriptions with textual content im-

Table 3: Official results for the Subtask 2b.

Team	Rank	M-F1	m-F1
LMEME	1	81.030	82.500
SuteAlbastre	2	80.964	83.500
DUTIR938	3	80.910	83.667
BCAmirs	4	80.337	82.500
Snarci	5	79.860	82.667
	...		
UMUTeam	8	**78.660**	**80.667**
	...		
Baseline	-	25.000	33.333

proves overall performance in a multimodal setting. This approach does not impose any restrictions on embedding images and text together in the same vector space when fine-tuning or training persuasion classification techniques. Rather, we merge the text from the meme with its description and use this combined dataset as input for fine-tuning the pre-trained transformer-based model.

6 Error analysis

As far as we know, the organizers did not provide the gold labels of the test set to the participants. Therefore, we did a bug analysis based on the results of the development set.

Table 4 shows the results and the ranking we got with our development set approach.

In subtask 1 our approach is based on a fine-tuned model of RoBERTa-large. We obtained an H-F1 of 62.201 and an M-F1 of 35.514. From the confusion matrices (see Figure 5), we can see that the model didn't correctly predict any instances of the classes `Misrepresentation of Someone's Position (Straw Man)`, `Obfuscation, Intentional vagueness, Confusion`, `Presenting Irrelevant Data (Red Herring)`, and `Reductio ad Hitlerum`, indicating a possible class imbalance or lack of representative features for these classes. In addition, the F1 score of the `Whataboutism` and `Causal Oversimplification` class is relatively low compared to other classes, suggesting that the model has difficulty correctly identifying instances of this class, possibly due to ambiguous or overlapping features with other classes.

Subtask 2a is a hierarchical multi-label classification problem, but unlike Subtask 1, it uses a multimodal dataset, i.e. it uses both textual and visual multimodal features to identify 22 persuasion techniques. In this case, our model achieved an H-F1 of 67.902 and an M-F1 of 36.841, which is an improvement over the unimodal approach (Subtask 1). However, similar to the model in Subtask 1 (see Figure 6), it failed to predict any instances of the classes `Misrepresentation of Someone's Position (Straw Man)`, `Obfuscation, Intentional Vagueness, Confusion`, and `Presenting Irrelevant Data (Red Herring)`, and it obtained a lower F1 score in `Casual Oversimplification` and `Appeal to (Strong) Emotions`, except for the class `Reductio ad Hitlerum`, for which it correctly predicted 2 instances. This could be due to insufficient training data or ineffective feature representation for these classes.

Regarding Subtask 2b, a multimodal binary classification problem, our approach achieved an M-F1 of 76.836, and in Figure 7 we can see that the model misclassified 40% of the examples as non-propagandistic and 9% as propagandistic.

7 Conclusion

In this paper, we describe our participation in a SemEval task focused on identifying persuasive techniques in memes using a multimodal approach. For all three subtasks, we used the fine-tuning approach with the RoBERTa-large model for text features and LlaVa to extract image descriptions and combine them with the meme text. Our system achieved the tenth best result with an H-F1 of 64.774%, the fourth best in Subtask 2a with an H-F1 of 69.003%, and the eighth best in Subtask 2b with a macro-F1 of 78.660%.

As further work, we will evaluate the relationship between the persuasion techniques used in the different domains evaluated by our team. In this sense, we propose to re-annotate the Spanish SatiCorpus 2021 (García-Díaz and Valencia-García, 2022) and the PoliticES 2022 dataset (García-Díaz et al., 2022), which are focus on figurative language and politics respectively, with the 22 persuasion techniques and evaluate the reliability of using binary and hierarchical multi-label classification approaches. Another area where persuasion techniques may be present is in the identification of misogyny (García-Díaz et al., 2023).

Table 4: Results for dev split.

-	Rank	H-F1	H-P	H-R	M-F1	m-F1
Subtask 1	14	62.201	71.111	55.276	35.514	52.439
Subtask 2a	4	67.902	75.151	61.929	36.841	57.124
Subtask 2b	11	-	-	-	76.836	80.667

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way to make Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, this work was funded by the Spanish Government, the Spanish Ministry of Economy and Digital Transformation through the Digital Transformation through the "Recovery, Transformation and Resilience Plan" and also funded by the European Union NextGenerationEU/PRTR through the research project 2021/C005/00149877. Mr. Ronghao Pan is supported by the "Programa Investigo" grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

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.

Emilio Ferrara, Stefano Cresci, and Luca Luceri. 2020. Misinformation, manipulation, and abuse on social media in the era of covid-19. *Journal of Computational Social Science*, 3:271–277.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

José Antonio García-Díaz, Salud M Jiménez Zafra, María Teresa Martín Valdivia, Francisco García-Sánchez, Luis Alfonso Ureña López, and Rafael Valencia García. 2022. Overview of politices 2022: Spanish author profiling for political ideology. *Procesamiento del Lenguje Natural*.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2023. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, 9(3):2893–2914.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Połap, Thippa Reddy Gadekallu, and Zunera Jalil. 2023. Multimodal religiously hateful social media memes classification based on textual and image data. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

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 *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7-9, 2006. Proceedings 19*, pages 395–406. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

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. *CoRR*, abs/1907.11692.

Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. 2023. Overview of dipromats 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 71:397–407.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132.

A Confusion matrices for the error analysis with the test set

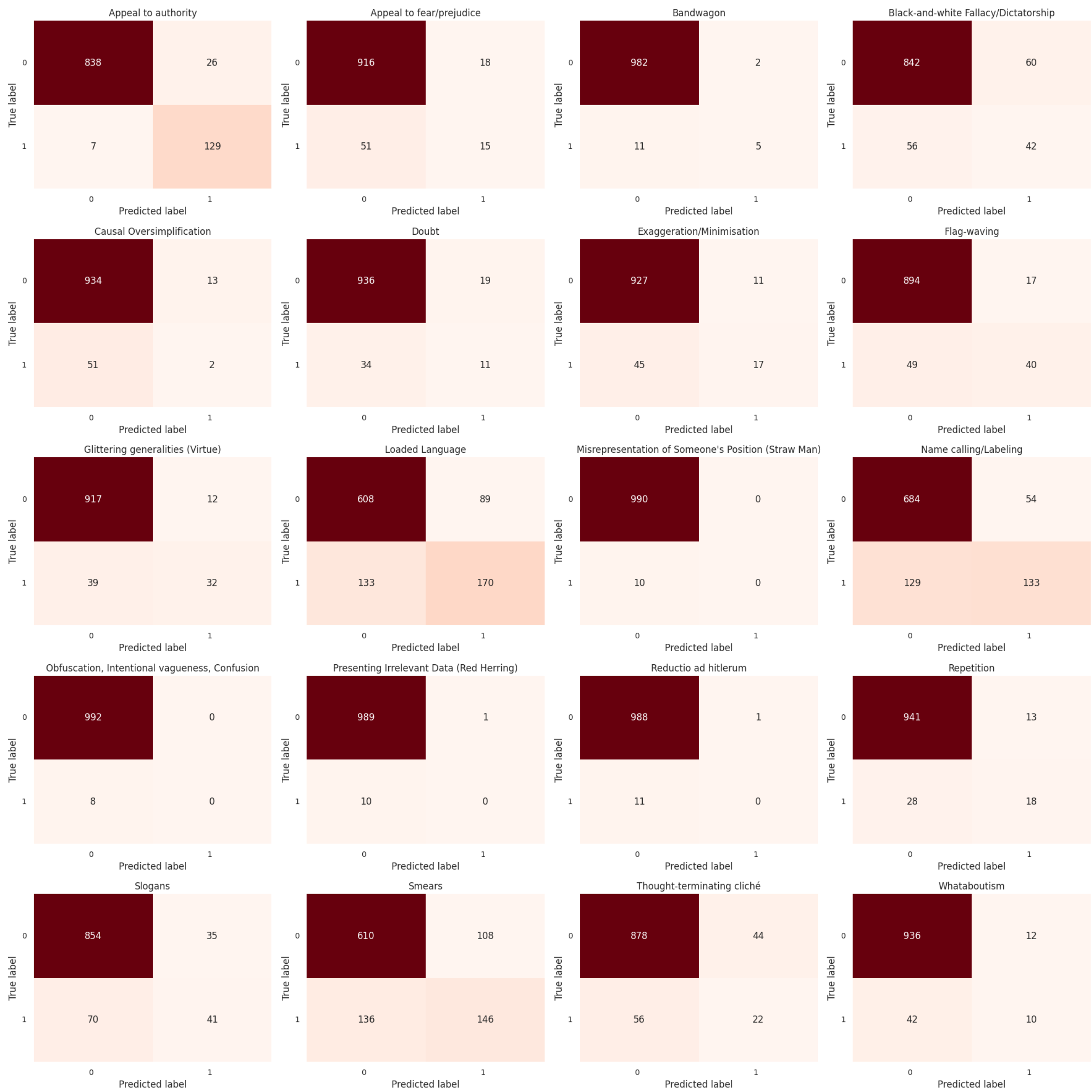

Figure 5: The confusion matrix of the model in the dev set of subtask 1.

Figure 6: The confusion matrix of the model in the dev set of subtask 2a.

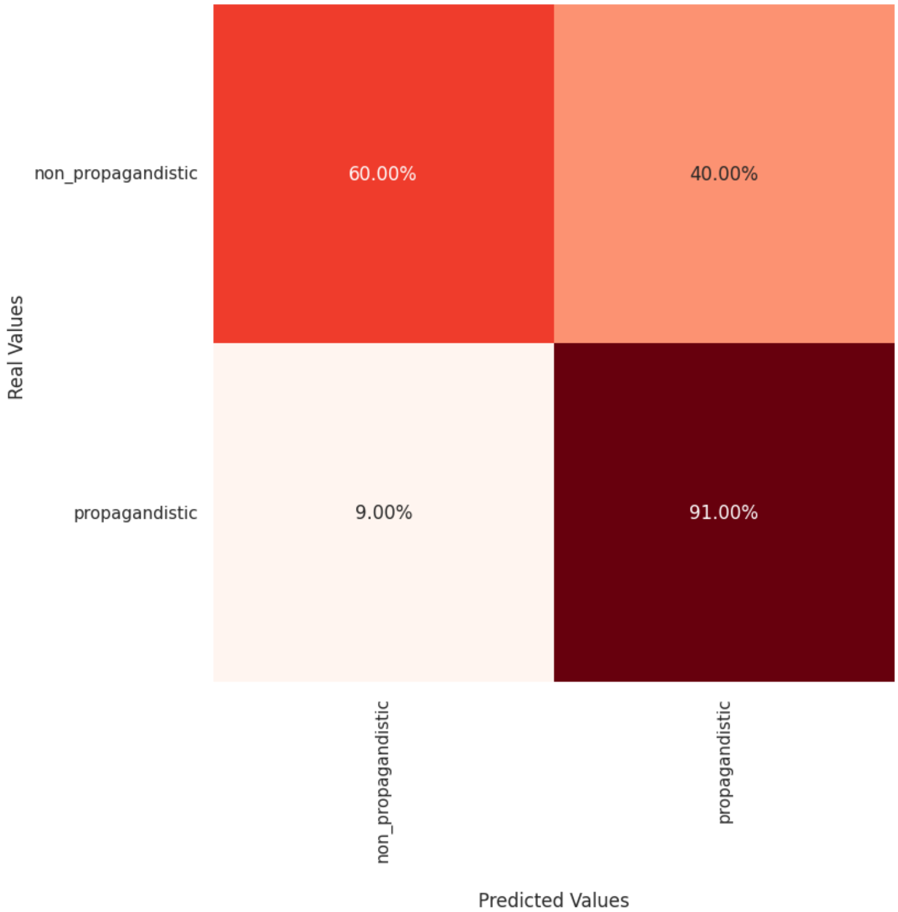

Figure 7: The confusion matrix of the model in the dev set of subtask 2b.

B Classification report with the dev set.

Table 5: Classification report of subtask 1 in the dev set.

	precision	**recall**	**f1-score**
Appeal to authority	83.2258	94.8529	88.6598
Appeal to fear/prejudice	45.4545	22.7273	30.3030
Bandwagon	71.4286	31.2500	43.4783
Black-and-white Fallacy/Dictatorship	41.1765	42.8571	42.0000
Causal Oversimplification	13.3333	03.7736	05.8824
Doubt	36.6667	24.4444	29.3333
Exaggeration/Minimisation	60.7143	27.4194	37.7778
Flag-waving	70.1754	44.9438	54.7945
Glittering generalities (Virtue)	72.7273	45.0704	55.6522
Loaded Language	65.6371	56.1056	60.4982
Misrepresentation of Someone's Position (Straw Man)	00.0000	00.0000	00.0000
Name calling/Labeling	71.1230	50.7634	59.2428
Obfuscation, Intentional vagueness, Confusion	00.0000	00.0000	00.0000
Presenting Irrelevant Data (Red Herring)	00.0000	00.0000	00.0000
Reductio ad hitlerum	00.0000	00.0000	00.0000
Repetition	58.0645	39.1304	46.7532
Slogans	53.9474	36.9369	43.8503
Smears	57.4803	51.7730	54.4776
Thought-terminating cliché	33.3333	28.2051	30.5556
Whataboutism	45.4545	19.2308	27.0270
micro avg	60.8918	46.0475	52.4394
macro avg	43.9971	30.9742	35.5143
weighted avg	58.2540	46.0475	50.6873
samples avg	46.3050	38.2436	39.3535

Table 6: Classification report of subtask 2a in the dev set.

	precision	recall	f1-score
Appeal to (Strong) Emotions	33.3333	17.8571	23.2558
Appeal to authority	83.1250	93.0070	87.7888
Appeal to fear/prejudice	46.4286	16.6667	24.5283
Bandwagon	75.0000	16.6667	27.2727
Black-and-white Fallacy/Dictatorship	38.6555	44.6602	41.4414
Causal Oversimplification	25.0000	03.5714	06.2500
Doubt	41.9355	25.0000	31.3253
Exaggeration/Minimisation	57.8947	16.1765	25.2874
Flag-waving	66.0000	53.6585	59.1928
Glittering generalities (Virtue)	63.0769	44.5652	52.2293
Loaded Language	70.7031	59.1503	64.4128
Misrepresentation of Someone's Position (Straw ...	00.0000	00.0000	00.0000
Name calling/Labeling	71.3592	56.3218	62.9550
Obfuscation, Intentional vagueness, Confusion	00.0000	00.0000	00.0000
Presenting Irrelevant Data (Red Herring)	00.0000	00.0000	00.0000
Reductio ad hitlerum	50.0000	12.5000	20.0000
Repetition	60.6061	43.4783	50.6329
Slogans	58.3333	42.6087	49.2462
Smears	72.7099	75.5952	74.1245
Thought-terminating cliché	28.7879	24.3590	26.3889
Transfer	59.3909	42.7007	49.6815
Whataboutism	60.0000	24.1935	34.4828
micro avg	64.7779	51.0870	57.1236
macro avg	48.2882	32.3971	36.8408
weighted avg	61.8299	51.0870	54.7241
samples avg	62.8705	52.9558	54.3457

Table 7: Classification report of subtask 2b in the dev set.

	precision	recall	f1-score
non_propagandistic	76.9231	60.0000	67.4157
propagandistic	81.9820	91.0000	86.2559
accuracy	80.6667	80.6667	80.6667
macro avg	79.4525	75.5000	76.8358
weighted avg	80.2957	80.6667	79.9759