LeCun at SemEval-2021 Task 6: Detecting Persuasion Techniques in Text Using Ensembled Pretrained Transformers and Data Augmentation

Dia Abujaber* **Ahmed Qarqaz*** **Malak Abdullah**
Jordan University of Science and Technology
Irbid, Jordan
daabujaber17, afalqarqaz17@cit.just.edu.jo
mabdullah@just.edu.jo

Abstract

This paper presents one of the top systems for the SemEval-2021 task 6 (Dimitrov et al., 2021), "detection of persuasion techniques in text and images". The proposed system, LeCun, targets subtask-1 for detecting propaganda techniques based on the textual content of a meme. We have used an external dataset from a previous relevant SemEval competition (Martino et al., 2020). We also have articulated another dataset using data-augmentation techniques. The final proposed model consisted of 5 ensemble transformers (four RoBERTa models and one DeBERTa), each trained on either a different dataset or preprocessing. Apparently, ensembling models trained on different datasets improve performance more than ensembling models trained on the same dataset/preprocessing. Also, it is obvious, fine-tuning the model on the Competition dataset after training it for a few epochs on the other datasets would improve the f1-micro up to 0.1 additional scores. The final model achieved an f1-micro score of 0.512 on the test dataset and an f1-micro of 0.647 on the development dataset.

1 Introduction

The definition of Memes was constantly changing since it was first conceived. But, Memes eventually got an academic definition, called an "Internet Meme". As Davison (2012) Internet Meme can roughly be defined as "a piece of culture, typically a joke, which gains influence through online transmission". But what makes Internet memes unique is the speed of their transmission and the fidelity of their form. Therefore the Internet meme can act as a powerful medium for persuasion techniques that preach an ideology or way of thinking. (Moody-Ramirez and Church, 2019)

On the other hand, the term "propaganda" is defined as a form of communication that employs persuasive strategies and attempts to achieve a response that furthers the desired intent of the propagandist (Jowett and O'donnell, 2018). With the rise of social media, a new form of propaganda rises called "Computational Propaganda." The author in (Woolley and Howard, 2017) defined Computational Propaganda as "The use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks".

Task 6 at SemEval-2021 (Dimitrov et al., 2021), detection of persuasion techniques in text and images, defined three subtasks. The first two subtasks deal with the textual contents of memes that ask the participants to identify which of the 20 propaganda techniques are in the text. While the third subtask encourages the participants to determine which of the 22 techniques are in the meme's textual and visual content. This paper proposes a solution for subtask1 that uses pre-trained language models to detect propaganda and possibly even identify the persuasion strategy that the propaganda sample employs.

The rest of the paper is broken down as follows. Section 2 discusses related-work to the task of propaganda identification. Section 3 provides a description of the data and the pre-processing techniques used. Section 4 describes the proposed system and architecture. Section 5 presents system analysis. Finally, the conclusion and future work are provided in Section 6.

2 Related Work

There have been efforts in persuasion techniques identification and classification using machine and deep learning-based approaches. The authors in (Al-Omari et al., 2019) used word embeddings with BERT (Devlin et al., 2019) and BiLSTM (Schuster

* Equal Contribution

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1068–1074
Bangkok, Thailand (online), August 5–6, 2021.

and Paliwal, 1997) for binary detection of propaganda spans. Authors in (Altiti et al., 2020) experimented with a CNN (LeCun et al., 1999), BiLSTM and BERT and showed BERT to have the best accuracy on classifying persuasion techniques in propaganda spans. Also, the authors in (Jurkiewicz et al., 2020) used a RoBERTa model (Liu et al., 2019), a class-dependent re-weighting method and used a semi-supervised learning technique of self-training and demonstrated the effects of these techniques in an ablation study. A group of researchers (Morio et al., 2020) experimented with a variety of PLMs (pre-trained language models), including BERT, GPT-2 (Radford et al., 2019), RoBERTa, XLM-RoBERTa (Conneau et al., 2019), XLNet (Yang et al., 2019) and XLM (CONNEAU and Lample, 2019). And have demonstrated that RoBERTa and XLNet generally perform better for propaganda detection.

3 Data Description

In this section, we describe the data and the task and the preprocessing step

3.1 Data

The dataset used during our experiments has been provided by the SemEval-2021 Task 6 (Dimitrov et al., 2021). The dataset, "Competition dataset", consists of short text samples that were extracted from Memes. We have also resorted to using an external dataset (Da San Martino et al., 2019) that is comprised of news articles with propaganda spans, "External Dataset". To use the External dataset effectively, we needed to chop down the news articles closer to the text's length in the current dataset and take only the text fragment that contained the propaganda and the corresponding label representing the propaganda technique in that text fragment.

3.2 Data Preprocessing

Our data preprocessing pipeline consists of two components, 1) Data cleaning 2) Data augmentation. In this section, we will describe the techniques we used in each component.

3.2.1 Data Cleaning

To increase performance accuracy, some data preprocessing techniques have been tested. We have experimented with typical pre-processing techniques, such as "Stop-Words Removal", which refers to removing commonly used words (such as "the", "a", "an", "in") to eliminate noise that may otherwise hinder the model's ability to learn and predict sequences. We have also experimented with "Stemming" which refers to the process of reducing inflection in words (e.g., connect, connected, connection) to their root form (e.g., connect). The specific Stemming algorithm that was used is Porters Algorithm (Porter, 1980).

3.2.2 Data Augmentation

We experimented with Data Augmentation (Wei and Zou, 2019). This is the process of using the original given data to produce more data to increase the given dataset size. Data Augmentation has been proved to be useful when dealing with small datasets. Although this technique is more prevalent in computer vision tasks, there are some versions of the technique that are specifically tailored to work with text data as described at (Wei and Zou, 2019). These techniques include Synonym Replacement, Random Insertion, Random Swap, Random Deletion, Back-translation. Table 1 shows examples on generating data using data-augmentation. This was done by using the "Easy Data Augmentation" library (Wei and Zou, 2019).

Back-translation was only applied on the Competition dataset and the other four techniques on the External dataset. For each sentence in the External dataset, 0.1 percent of Synonym Replacement, Random Insertion, Random Swap, and Random deletion was applied. We did that nine times per sentence for each sample. In the Back-translation, AWS API was used to translate the text from English to Dutch back to English. Also, from English to Dutch to Russian, back to English. For each sample, we generate two additional samples. The Competition dataset has a size of 487. After merging the External dataset and the Competition dataset, we ended up with a dataset of size 18,571. This data will be referred to as the "Competition + External Dataset". After applying data augmentation on the Competition + External dataset, we ended up with a dataset of size 52,966. This data will be referred to as the "Augmented Dataset."

4 System Description

Different model architectures have experimented with different pre-processing techniques. The final system ended up ensembling five models, each

See `https://github.com/jasonwei20/eda_nlp` for the data augmentation code

Original	*This paper will describe our system in detecting propaganda in memes*
Synonym Replacement	*This theme will describe our arrangement in detecting propaganda in memes*
Random Insertion	*This key out paper will describe our system meme in detecting propaganda in memes*
Random Swap	*This paper in describe our system in detecting propaganda will memes*
Random Deletion	*This paper will describe system in detecting propaganda in memes*
Back-translation	*This document describes our Memorandum Propaganda Detection System*

Table 1: Generated Samples using Data-augmentation

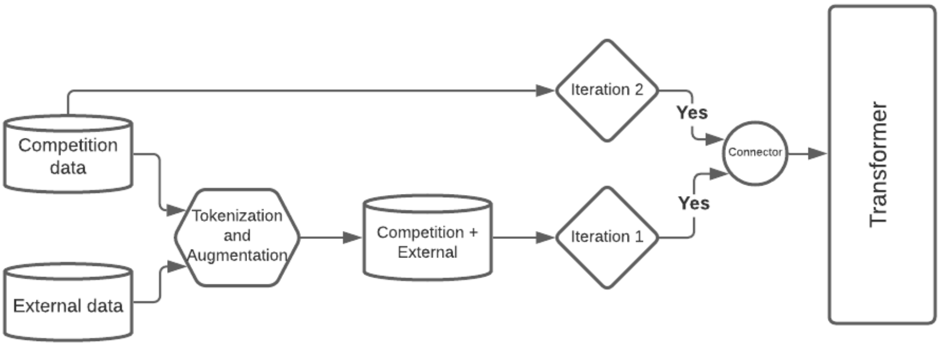


Figure 1: Approach 2 Training Schema

trained on a different approach. The ensemble model consists of *1 DeBERTa (He et al., 2021) model* and *4 RoBERTa models* each trained in a different approach or data pre-processing technique. We have two training approaches we used in training our models. The first one is a typical fine-tuning. The second approach consists of two iterations. In the first iteration, the model is trained on the pre-processed dataset. In the second iteration, the model from the first iteration is fine-tuned on the Competition dataset exclusively. Figure 1 demonstrates the second approach.

4.1 Proposed System

The system is an ensemble model of 5 classifiers. One of them is using the DeBERTa large classifier, and the rest are RoBERTa large classifiers. Each classifier is trained on a different approach/pre-processing. For the DeBERTa large classify, the Augmented dataset was used with the stop words removed and lowered text case. Then we trained it on the first approach for six epochs. It achieved F1 micro of 0.554 on the development set. As mentioned earlier, there are 4 RoBERTa large classifiers; the first classifier is trained on the Competition dataset and the External dataset without augmentation. We dropped samples that do not have any propaganda technique and trained the model on the first approach for four epochs. It achieved an F1 micro score of 0.550. The second RoBERTa classifier has the same pre-processing as the first RoBERTa classifier but is trained on the second approach. It achieved an F1 micro score of 0.602 on the development set. The third RoBERTa classifier is trained on the Augmented dataset with the stop words and trained using the first approach with four epochs. It achieved F1 micro of 0.54. The fourth RoBERTa classifier is the same as the third model but fine-tuned on the Competition dataset and achieved an f1 score of 0.62. Table 2 summarizes the performance of the classifiers of LeCun's ensemble model.

5 System Analysis

5.1 Ensemble Analysis

This section will be analyzing different combinations of the models that lead to the final proposed system. In the second approach, we noticed that fine-tuning the classifiers from the first iteration on the Competition dataset will always boost the performance up to 0.1 additional f1 micro scores on the development set. However, when it came to the ensemble model, it turned out that the ensemble model with classifiers from the second training approach doesn't increase the overall performance, and sometimes it decreased it. Table 3 demonstrates the performance of different classifiers combinations. Ensemble (A) consists of classifier (3) with f1 micro of 0.602 and classifier (5) with f1 micro of 0.62. After the ensemble, the overall score

#	Model Type	Epochs	LR	SQM	BS	Dataset	Approach	F1-Macro	F1-Micro
1	DeBERTa Large	3	2e-06	164	8	Augmented	1	0.430	0.554
2	RoBERTa Large	2	2e-05	64	8	Competition + External	1	0.340	0.550
3	RoBERTa Large	3	2e-05	64	8	Competition + External	2	0.388	0.602
4	RoBERTa Large	4	2e-05	128	16	Augmented	1	0.358	0.540
5	RoBERTa Large	2	2e-05	128	16	Augmented	2	0.430	0.622

Table 2: Models Hyper-parmeters. LR (Learning Rate), SQM (Sequence Max Length), BS (Batch Size)

	Ensemble Combination	F1 Micro	F1 Macro
A	(3)(5)	0.58	0.40
B	(4)(5)	0.59	0.40
C	(3)(4)(5)	0.62	0.41
D	(2)(3)(4)(5)	0.60	0.37
E	(1)(2)(3)	0.56	0.39
F	(1)(2)(3)(4)	0.59	0.40
G	(1)(2)(3)(4)(5)	0.64	0.44

Table 3: Performance on Different Ensemble Combinations

dropped to 0.58. However, Ensemble (A) suggests that ensembling classifier (3) and classifier (5) isn't optimal. In the final ensemble model (G), we noticed that the overall score would decrease if we removed one of these models. For example, in ensemble (F), classifier (5) was dropped, and both f1 micro and macro decreased.

5.2 Error Analysis

This section examines the ensemble model weaknesses to give insight on what to do next to improve the model performance. We have generated the confusion matrix and scores for each class for the test set (see Appendix). In the confusion matrices, the "None" class indicates that either the model predicted an incorrect class (not in the ground-truth labels set) or it didn't predict the correct class (in the ground-truth labels set but not in the predicted labels set). It is worth noting that the correctly classified "None" (not in the predicted labels set and not in the ground-truth labels set) is not provided in the model evaluation confusion matrix. We noticed that the model performs poorly at detecting the classes (last column) in the input test (see Figure A2). One possible explanation for this is that the model is trained on data that has a lot of samples without propaganda (count of labels = 0) (see Figure A1). In addition to that, the label matrix is sparse (zero is the dominant label in a one-hot vector). One possible solution is to remove samples with zero labels and rely on the sparsity of vectors in detecting samples without propaganda. Another possible solution is to train a Two-Stage model, where the first stage filters out non-propaganda samples and the second stage classify the propaganda samples.

6 Conclusion and Future Work

In this paper, we presented our proposed system, LeCun, for detecting propaganda in contextual content. We have used an external dataset from a previous SemEval competition and performed a data-augmentation on the external dataset to expand the dataset size. We have also investigated different ensemble combinations for state-of-the-art pre-trained language models. However, there are many questions we got throughout our participation in this competition which made us curious to investigate. These questions are:- What is the influence of the data augmentation on the model performance? What is the influence of using an external dataset? How can the model weaknesses be improved? How can span identification help in improving the score of technique classification?

For future work, we will be working on answering these questions. We plan to do more in-depth experimenting with different augmentation techniques and different model architectures. We will also investigate the influence of the external dataset by training models on the competition dataset, external dataset separately and compare the final results of each.

References

Hani Al-Omari, Malak Abdullah, Ola AlTiti, and Samira Shaikh. 2019. JUSTDeep at NLP4IF 2019 task 1: Propaganda detection using ensemble deep learning models. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 113–118, Hong Kong, China. Association for Computational Linguistics.

Ola Altiti, Malak Abdullah, and Rasha Obiedat. 2020. JUST at SemEval-2020 task 11: Detecting propaganda techniques using BERT pre-trained model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1749–1755, Barcelona (online). International Committee for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China.

Patrick Davison. 2012. The language of internet memes. *The social media reader*, pages 120–134.

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.

Dimiter Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention.

Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage publications.

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

Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. 1999. *Object Recognition with Gradient-Based Learning*, pages 319–345. Springer Berlin Heidelberg, Berlin, Heidelberg.

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.

G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles.

Mia Moody-Ramirez and Andrew B Church. 2019. Analysis of Facebook meme groups used during the 2016 US presidential election. *Social Media+ Society*, 5(1):2056305118808799.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Samuel C Woolley and Philip Howard. 2017. Computational propaganda worldwide: Executive summary.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Appendix

Label	F1-score	Precision	Recall	Support
Appeal to authority	0.0	0.0	0.0	7
Appeal to fear/prejudice	0.57	0.40	0.47	10
Black-and-white Fallacy/Dictatorship	1.0	0.29	0.44	77
Causal Oversimplification	0.0	0.0	0.0	3
Exaggeration/Minimisation	0.60	0.47	0.53	19
Flag-waving	0.40	0.33	0.36	6
Name calling/Labeling	0.64	0.55	0.59	53
Loaded Language	0.80	0.74	0.77	100
Presenting Irrelevant Data (Red Herring)	0.0	0.0	0.0	4
Reductio ad hitlerum	1.0	0.33	0.50	3
Misrepresentation of Someone's Position (Straw Man)	0.0	0.0	0.0	1
Thought-terminating cliché	0.0	0.0	0.0	6
Bandwagon	0.0	0.0	0.0	1
Doubt	0.67	0.21	0.32	28
Repetition	0.0	0.0	0.0	1
Slogans	0.2	0.05	0.08	19
Whataboutism	1.0	0.1	0.18	10
Smears	0.67	0.18	0.28	45
Glittering generalities (Virtue)	0.0	0.0	0.0	11
Obfuscation, Intentional vagueness, Confusion	0.0	0.0	0.0	1

Table 1: Classification report of the submitted system on the test set

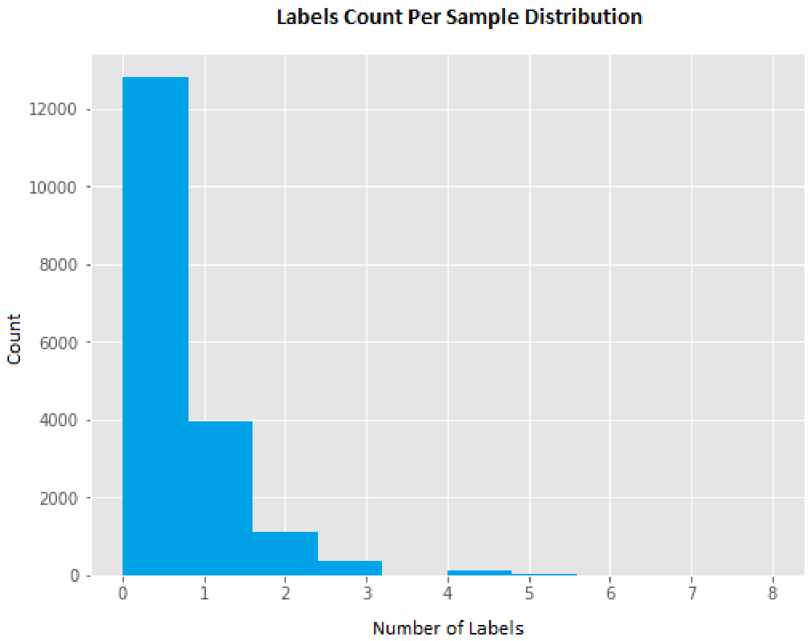
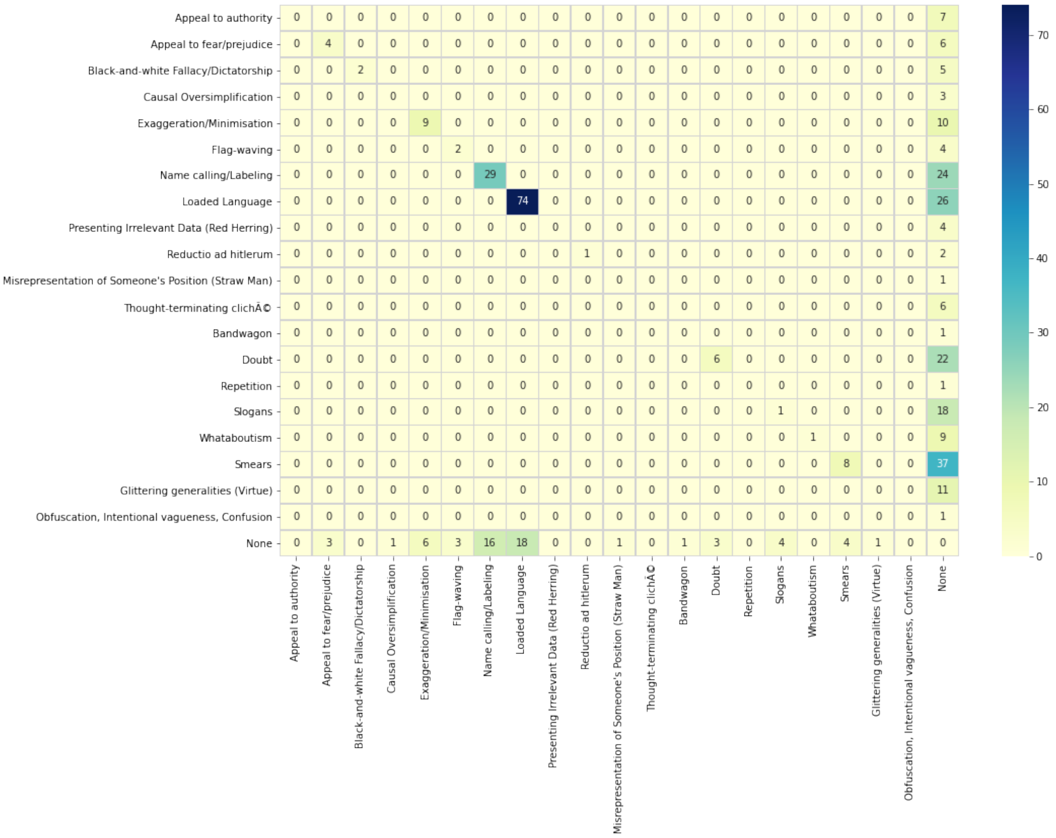


Figure A1: Labels count per sample distribution of Competition + official dataset



	Appeal to authority	Appeal to fear/prejudice	Black-and-white Fallacy/Dictatorship	Causal Oversimplification	Exaggeration/Minimisation	Flag-waving	Name calling/Labeling	Loaded Language	Presenting Irrelevant Data (Red Herring)	Reductio ad hitlerum	Misrepresentation of Someone's Position (Straw Man)	Thought-terminating clichÃ©	Bandwagon	Doubt	Repetition	Slogans	Whataboutism	Smears	Glittering generalities (Virtue)	Obfuscation, Intentional vagueness, Confusion	None
Appeal to authority	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
Appeal to fear/prejudice	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
Black-and-white Fallacy/Dictatorship	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
Causal Oversimplification	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
Exaggeration/Minimisation	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
Flag-waving	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Name calling/Labeling	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	24
Loaded Language	0	0	0	0	0	0	0	74	0	0	0	0	0	0	0	0	0	0	0	0	26
Presenting Irrelevant Data (Red Herring)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Reductio ad hitlerum	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
Misrepresentation of Someone's Position (Straw Man)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Thought-terminating clichÃ©	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
Bandwagon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Doubt	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	22
Repetition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Slogans	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	18
Whataboutism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	9
Smears	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	37
Glittering generalities (Virtue)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Obfuscation, Intentional vagueness, Confusion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
None	0	3	0	1	6	3	16	18	0	0	1	0	1	3	0	4	0	4	1	0	0

Figure A2: Confusion matrix of the submitted system on the test set - i-th row and j-th column entry indicates the number of samples with true label being i-th class and predicted label being j-th class