NLUBot101 at SemEval-2023 Task 3: An Augmented Multilingual NLI Approach Towards Online News Persuasion Techniques Detection

Genglin Liu, Yi R. Fung, Heng Ji
University of Illinois Urbana-Champaign
{genglin2, yifung2, hengji}@illinois.edu

Abstract

We describe our submission to SemEval 2023 Task 3, specifically the subtask on persuasion technique detection. In this work, our team tackled a novel task of classifying persuasion techniques in online news articles at a paragraph level. The low-resource multilingual datasets, along with the imbalanced label distribution, make this task challenging. Our team presented a cross-lingual data augmentation approach and leveraged a recently proposed multilingual natural language inference model to address these challenges. Our solution achieves the highest macro-F1 score for the English task, and top 5 micro-F1 scores on both the English and Russian leaderboards. We have made the source code of our models and experiments publically available at [1].

1 Introduction

We describe UIUC BLENDER Lab's participation in SemEval 2023 Task 3. Detecting persuasion techniques in multilingual online news articles is valuable for several reasons. The rise of digital media and social networks has led to an increase in the amount of news and information available, making it difficult for individuals to navigate the vast amount of information and identify credible sources. By detecting persuasion techniques in news articles, NLP can help individuals distinguish between objective reporting and biased reporting, become more aware of these techniques and make more informed decisions. From a journalist's perspective, detecting persuasion techniques in online news articles using an automated system can help them and editors identify and remove any biased language, leading to more objective reporting and higher-quality journalism. Regarding detecting persuasion techniques in online news articles, it is useful in a multilingual setting for several reasons.

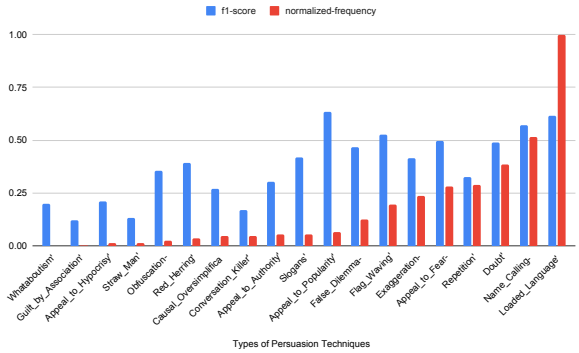

Figure 1: F1-score on English Dev Set and Normalized Frequency for Each Label

In today's interconnected world, multilingual NLP can help identify the persuasion techniques used in news articles in different languages to understand the global media landscape better. Multilingual NLP can help identify patterns and common techniques used to spread propaganda and misinformation in different languages and help combat these issues.

SemEval 2023 released a new dataset covering several aspects of what makes a text passage persuasive. Our training data comes from articles in six languages: English, French, German, Italian, Polish, and Russian. They were collected from 2020 to the middle of 2022, focusing on a set range of topics. The details are presented by the task organizers (Piskorski et al., 2023). The task provides news articles in six languages, and our team decided to focus on the persuasion technique detection subtask since it is unique, challenging, and aligns well with our research interests. We find that our system was able to generate reasonable results across the 6 languages presented by the task organizers, as well as generalize on unseen languages such as Greek and Spanish.

[1] https://github.com/yrF1/SemEval23-Task-3-UIUC-Team/tree/main

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1636–1643

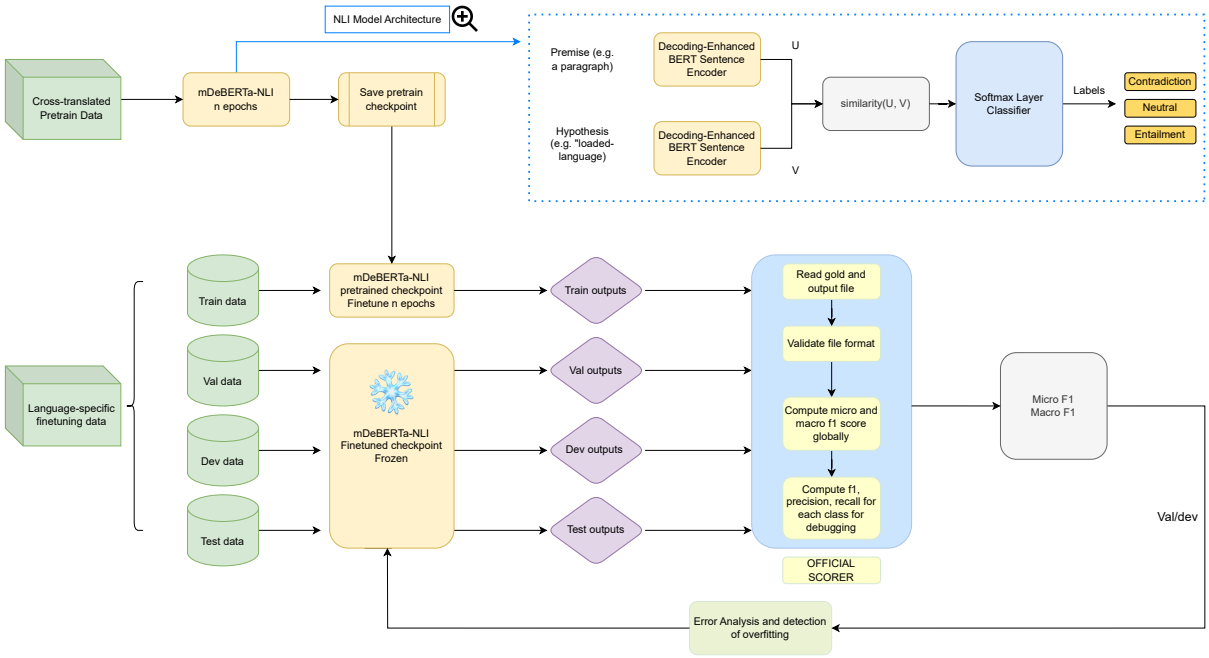

Figure 2: System Architecture and Training pipeline for our NLI based propaganda detection system

2 Background

2.1 Problem Formulation

Given a news article, our task is to identify the persuasion techniques in each paragraph. This is a multi-label task at the paragraph level. News articles are given in the exact same format but in a multilingual setup. Training and development data sets are provided in 6 languages, and our system is later evaluated on 3 additional "surprise" languages at test time. As depicted in Figure 1, there are 19 persuasion technique labels in the English dataset, and 23 total for all languages. As we can see in both Figure 1 and Table 1, the main bottlenecks for this task are (1) scarcity of the input data, (2) highly imbalanced label distribution, and (3) some labels are semantically similar and hard to differentiate. The model performance tends to suffer from low-frequency labels despite our effort to alleviate the issue through extensive data augmentation. To address these issues, we leverage all the training articles that are provided and use a strong multilingual Natural language Inference model to perform classification on each segment of a given text.

2.2 Related Work

There have been two related tasks in previous years of the SemEval workshop. SemEval 2020 Task-11 (Martino et al., 2020) was the first one that proposed a shared task regarding propaganda detection on news articles. A year later, SemEval 2021 hosted task 6: detecting persuasion techniques in a multi-modal setting (Dimitrov et al., 2021). We inspected the dataset provided in both of the two previous SemEval tasks, and realized that the 2020 task does not have multilingual resources, and the 2021 task only contains a limited amount of text data. Besides relevant tasks from the workshop in the past, there have been a number of other studies that are relevant to persuasion technique detection, or misinformation detection news/media analysis in general (Fung et al., 2022; Reddy et al., 2021; Pöyhönen et al., 2022). We also found inspiration in using cross-lingual machine translation to boost NLP performance (Whitehead et al., 2020) as well as leveraging other data augmentation systems that generate propaganda-loaded text (Huang et al., 2022).

SubTask3	EN	FR	GE	IT	PO	RU
Train	446	158	132	227	145	143
Dev	90	53	45	76	49	48

Table 1: Dataset train/dev split across 6 languages for subtask 3

3 System Overview

The two arguably most essential aspects of solving a machine learning problem are the data and the model. In this section, we discuss how training data is prepared and how we selected our model.

3.1 Baseline Models

We initially tried a naive BERT model (Devlin et al., 2018) with a classification head that predicts multi-label vectors based on a fixed threshold, but it did not outperform the official Support Vector Machines (SVMs) baseline, perhaps suffering from the limited amount of training samples and could not establish a good pattern from the dataset. We then used a BART-mNLI model (Lewis et al., 2019) and obtained reasonable performance on this subtask. We decided to use a Natural Language Inference (NLI) model to approach this problem because of two advantages that this type of model holds. First, it is arguably better at understanding language nuances: NLI models are designed to understand the nuances of natural language (Storks et al., 2019). This means they are better equipped to deal with language that is complex, ambiguous, and open to interpretation. Another advantage of NLI models is flexibility: they can be trained to classify text based on a wide range of criteria, and can transfer reasoning capability to new or related persuasion techniques categories.

3.2 mDeBERTa-v3 NLI Model Backbone

This model is suitable for multilingual classification since it can perform natural language inference on 100 different languages. DeBERTa is an advanced variant of the BERT model with disentangled attention and an enhanced masked encoder part (He et al., 2020). DeBERTa-v3 is a further improvement using a more sample-efficient pretraining task called replaced token detection (RTD) (He et al., 2021). Microsoft pre-trained the underlying mDeBERTa-v3-base model using the 100-language CC100 multilingual dataset (Conneau et al., 2019). Due to the multilingual setup of our task, we adopted a variant of mDeBERTa-v3 [2] that was refined on the multilingual-NLI-26lang-2mil7 (Parrish et al., 2021) and the XNLI dataset (Conneau et al., 2018). More than 2.7 million hypothesis-prediction pairs in 27 languages are found in these datasets, making the mDeBERTa-v3 model one the best-performing multilingual transformer model for our persuasion technique detection task.

[2] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

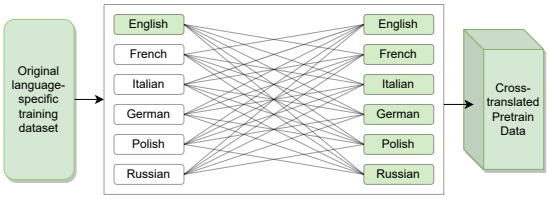

Figure 3: Constructing a pre-train dataset with cross-translated documents

3.3 Pretraining on Augmented Dataset

We use the development and test sets as provided by the task organizers. Concretely, the number of articles are shown in Table 1. Starting with the original development and training set, we pass these articles into a cross-translation module where the articles are cross translated into other languages as illustrated in Figure 3. For example, the 446 English articles will produce an equal number of articles in French, German, Italian, Polish, and Russian. And the same procedure is applied to every o ther language. After the augmentation step, we obtained a larger set of training and development data, specifically 1251 training articles and 361 dev articles in all languages.

We construct a "pretraining" dataset by combining all of the augmented articles in all languages. This pretraining dataset by our construction has 7506 articles. The development set is left alone to ensure no data leakage. In subtask 3, the machine learning model makes inference on each individual paragraph, so the overall amount of training/dev instances is more than the number of articles. We further isolated each persuasion technique label from the list of labels for each paragraph. For example, if a given paragraph has two distinct annotated labels, then we would end up having two copies of the same paragraph, each with one distinct label so that the model sees two instances and learns single-label prediction in the pretraining stage. We ended up having roughly 260,000 pretraining paragraph-label pairs.

4 Experimental Setup

4.1 System Pipeline and Configurations

We hold out 20% of the training set instances (i.e. paragraphs) and curate a validation set from these samples. Our overall system pipeline is illustrated in Figure 2. For the purpose of data augmentation, we tried both MarianMT neural translation model (Junczys-Dowmunt et al., 2018) as well as the GoogleTrans API. Google Translate relies on

making networked requests while executing the translation but obtains a higher accuracy. Every article is translated from its original language into the other 5 available languages specified by the task organizers. For this task, we use micro and macro-F1 scores as our main evaluation metric. Micro-F1 was the metric used for official scoring in this subtask, though it's worth noting that our system is more competitive in terms of macro-F1 scores.

In order to make our system replicable and the results reproducible, we also provided some configuration details in this section. The development and test set results are all produced by the mDeBERTa-v3 mNLI model; we used a learning rate of 1e-5 and batch size of 12 for experiments on all of the languages. Our pretraining checkpoint is saved after three epochs of training on the cross-translated pretraining dataset, and then for each individual language, we further finetune for 3-6 epochs depending on the validation performance and then saves a language-specific checkpoint after that phase. In terms of hardware, our experiments are done on single P100 or V100 Nvidia GPUs with 16G of graphic memory.

4.2 Augmented Labels for the NLI Model

We improved the 23 persuasion technique labels by using their expanded definitions instead of single words or phrases, based on the official annotation guideline. For example, "loaded_language" is defined as using words and phrases with strong emotional implications to influence and convince the audience. Our experiments on the development set showed as much as a 6% improvement in micro-F1 score when using these enhanced label definitions in conjunction with the NLI model, which takes a given paragraph as the premise and the hypothesis as the expanded label definition.

5 Results

Our system was able to achieve good results across the subtask 3 leaderboards beating baseline scores by a sizable margin. Our system NLUBot101 achieved 0.36058 in Micro-F1 and 0.19722 on Macro-F1 on the English leaderboard, finishing as number 5 out of 23 teams. On the Russian board, our same system ranked 4th out of 19 teams. We submitted test set evaluations on all 9 languages, and the complete results can be seen on Table 2.

5.1 Micro vs. macro-F1 Score Interpretation

The micro-F1 score is computed globally by counting the total number of predictions across all classes, unlike the macro-F1 where the score is computed for each class independently and then averaged across all classes. Consequently, micro-F1 gives more weight to larger classes, while macro-F1 treats all classes equally. Our system's macro-F1 performance is the highest among all teams, making it the most competitive candidate if an end-user wants good classification performance on a particular less common persuasion technique.

5.2 Error Analysis

In this section, we describe the error analysis we conducted on our NLI model. This serves as a crucial step in the development and evaluation of our models, as it helps us identify the strengths and weaknesses of the system and provides insight into areas for improvement.

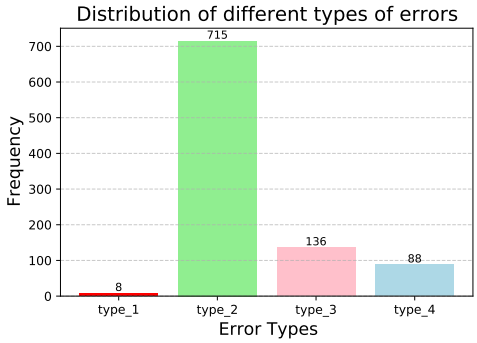

Figure 4: Distribution of different types of errors produced by our system in English

5.2.1 Definitions for 4 Different Error Types

We identify 4 distinct types of errors that occur in our model predictions between a set of predictions and groundtruth labels. The set relations can be described logically as follows:

- Type I error: $P \subset G$. Example: P = {label1}, G = {label1, label2}
- Type II error: $G \subset P$. Example: P = {label1, label2}, G = {label1}
- Type III error: $P \cap G \neq \emptyset$ and $P \not\subseteq G$ and $G \not\subseteq P$. Example: P = {label1, label2}, G = {label2, label3}
- Type IV error: $P \cap G = \emptyset$. Example: P = {label1, label2}, G = {label3}

	English	French	Italian	German	Polish	Russian	Spanish	Greek	Georgian
Official baseline micro-F1	0.19517	0.24014	0.39719	0.31667	0.17928	0.20722	0.24843	0.08831	0.13793
Our micro-F1 (rank/teams)	0.36058 (5/23)	0.39580 (6/20)	0.43506 (11/20)	0.42042 (7/20)	0.31970 (10/20)	0.32298 (4/19)	0.30459 (6/17)	0.15034 (10/16)	0.25362 (9/16)
Official baseline macro-F1	0.06925	0.09867	0.12152	0.08345	0.05932	0.08598	0.02007	0.00606	0.14083
Our macro-F1 (rank/teams)	**0.19722** (1/23)	0.25431 (5/20)	0.16371 (10/20)	0.17879 (8/20)	0.16911 (9/20)	0.20052 (3/19)	0.15092 (5/17)	0.09653 (9/16)	0.17235 (11/16)

Table 2: Official Leaderboard Results on the Test Sets

where P represents the prediction set and G represents the groundtruth set.

5.2.2 Failure Cases Analysis

We focus on results produced on the English development set for this section of analysis. Out of 3127 dev set instances, there are 1106 valid ones with non-empty predicted and groundtruth labels.

Type I error occurs when the model predicts a strict subset of the correct labels. This almost never happens in our model. One example is shown in Figure 5. The text says "NSA vacuuming up of a kazillion phone calls is, we're told, for our protection and well-being". The model predicts three labels, which are all included in the groundtruth, but missing one that is 'doubt'. This is hard to capture because the only phrase suggesting that label would be "we're told".

The model makes significantly more errors of type II than any other type as shown in Figure 4. Type II errors occur when the model's prediction includes all the correct labels, but it predicts more than it should. One of many examples is shown in Figure 5 of the Appendix section, where the model assigns two more labels than the groundtruth while there's no significant indication of them in the actual text. The model might behave this way if it assigns labels even with relatively low confidence. In that case, we speculate that this type of error could be alleviated by raising the entailment probability threshold in our NLI model.

Type III errors occur when the prediction and groundtruth share some but not all of the correct labels. This may be due to the model not capturing certain nuances in the data or the groundtruth labels being too broad. One example is provided in Figure 4. In this paragraph, the annotation includes 'appeal-to-fear' as a label but our model missed it and outputs 'loaded-language' instead of that. To reduce the number of type III errors, the model may need to have more fine-grained and label-specific guidance at training time to capture the nuances and subtlety in each persuasion technique.

Type IV errors occur when the prediction and groundtruth do not share any common labels. Here is an interesting failure case example for this type of error, illustrated in Figure 5. The original text reads "No doubt London's two-bob chancer of a Mayor Sadiq Khan would blame the absence of police on the streets on 'austerity' or the Tory cuts". The groundtruth only includes 'Appeal-to-hypocrisy', but our model predicts multiple different labels. We infer that the model might've included 'oversimplification' because of the phrase "no doubt", and included 'appeal-to-authority' because of the mention of "a mayor" and "police". Our system further predicts loaded-language and name-calling as labels for this text, and we find evidence through the phrase "two-bob chancer of a mayor" that would suggest these two persuasion techniques. We believe that the differences might come from the nuance of the text as well as the subjectivity of human annotations, but overall our model performs reasonably in these instances.

5.3 Performance on Individual Labels

During the development phase, we wanted to see a more detailed breakdown of how the model performs on each individual label. So we generated an additional Table 3 shown in the Appendix section due to the limited space. We make two major observations. First, the recall scores are significantly higher than the precision scores, which suggests that our model is capable of keeping the number of false negatives low, but might have a relatively higher number of false positive predictions. This observation agrees with our error analysis where we noticed that our model tends to predict "more than enough" while making inferences.

6 Conclusion

In this work, we described the submission of our team NLUBot101 in SemEval 2023 Task 3, specifically subtask 3 on multilingual persuasion technique detection at the paragraph level. We built the candidate based on a DeBERTa-v3 backbone and

then performed multilingual NLI on the task. Our model attains the highest macro-F1 performance in English among 23 teams, top 5 performance in English and Russian by micro-F1 score, and the top 10 in 8 out of the 9 languages that the model was evaluated on. We are excited about the performance of our model and the positive societal impact of such systems outside of NLP, but we also recognize its limitations through our failure analysis. For future work, we are interested in making our system more explainable and extending the model capability to more low-resource languages.

Acknowledgement

This research is based upon work supported in part by U.S. DARPA SemaFor Program No. HR001120C0123, DARPA INCAS Program No. HR001121C0165, and DARPA MIPS Program No. HR00112290105. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

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. In *Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: detection of persuasion techniques in texts and images. *arXiv preprint arXiv:2105.09284*.

Yi Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 28–34, Taipei. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does putting a linguist in the loop improve nlu data collection? *arXiv preprint arXiv:2104.07179*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Teemu Pöyhönen, Mika Hämäläinen, and Khalid Alnajjar. 2022. Multilingual persuasion detection: Video games as an invaluable data source for nlp. *arXiv preprint arXiv:2207.04453*.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogerio Feris, and Kate Saenko. 2020. Learning from lexical perturbations for consistent visual question answering. *arXiv preprint arXiv:2011.13406*.

A Appendix

Text ID	832931332-22	813452859-7	813552066-9	813953273-27
Original text	NSA vacuuming up of a kazillion phone calls is, we're told, for our protection and well-being.	Michael Swadling: I guess her only chance is if Labour decides that they want to dishonour democracy and effectively keep us in the EU.	Into this national crisis of epic proportions has just waded the clodhopping U.S. ambassador to Britain, billionaire Robert Wood 'Woody' Johnson.	No doubt London's two-bob chancer of a Mayor Sadiq Khan would blame the absence of police on the streets on 'austerity' or the 'Tory cuts'.
Predicted Labels	**Exaggeration-Minimisation, Flag_Waving, Loaded_Language**	**False_Dilemma-No_Choice**, Repetition, **Loaded_Language**, Name_Calling-Labeling, Doubt	**Name_Calling-Labeling, Exaggeration-Minimisation,** Loaded_Language	Causal_Oversimplification, Appeal_to_Authority, False_Dilemma-No_Choice, Repetition, Loaded_Language, Name_Calling-Labeling, Doubt
Groundtruth Label	Doubt, **Exaggeration-Minimisation, Flag_Waving, Loaded_Language**	**False_Dilemma-No_Choice, Loaded_Language**	Appeal_to_Fear-Prejudice, **Exaggeration-Minimisation, Name_Calling-Labeling**	Appeal_to_Hypocrisy
Error Type	Type I (under-predict)	Type II (over-predict)	Type III (limited intersection)	Type IV (disjoint)

Figure 5: Selected examples of Failure Cases in English. Bold text indicates common labels that appear in both the prediction and the groundtruth. Different highlights indicate parts of the original texts that suggest evidence for specific persuasion techniques.

Label	F1-score	precision	recall	frequency_of_label
Whataboutism	0.20	0.13	0.50	2
Guilt_by_Association	0.12	0.07	0.50	4
Appeal_to_Hypocrisy	0.21	0.18	0.25	8
Straw_Man	0.13	0.17	0.11	9
Obfuscation-Vagueness-Confusion	0.36	0.33	0.38	13
Red_Herring	0.39	0.29	0.63	19
Causal_Oversimplification	0.27	0.16	0.83	24
Conversation_Killer	0.17	0.10	0.64	25
Appeal_to_Authority	0.30	0.20	0.68	28
Slogans	0.42	0.28	0.86	28
Appeal_to_Popularity	0.63	0.73	0.56	34
False_Dilemma-No_Choice	0.47	0.32	0.89	63
Flag_Waving	0.53	0.36	0.97	96
Exaggeration-Minimisation	0.42	0.29	0.72	115
Appeal_to_Fear-Prejudice	0.50	0.35	0.84	137
Repetition	0.32	0.20	0.87	141
Doubt	0.49	0.37	0.73	187
Name_Calling-Labeling	0.57	0.41	0.94	250
Loaded_Language	0.62	0.45	0.99	483

Table 3: Dev Set Results on Per-Label F1 Scores and Label Frequency Distribution