LIIR at SemEval-2021 task 6: Detection of Persuasion Techniques In Texts and Images using CLIP features

Erfan Ghadery , Damien Sileo , and Marie-Francine Moens

Department of Computer Science (CS)
KU Leuven
{erfan.ghadery, damien.sileo, sien.moens}@kuleuven.be

Abstract

We describe our approach for SemEval-2021 task 6 on detection of persuasion techniques in multimodal content (memes). Our system combines pretrained multimodal models (CLIP) and chained classifiers. Also, we propose to enrich the data by a data augmentation technique. Our submission achieves a rank of 8/16 in terms of F1-micro and 9/16 with F1-macro on the test set.

1 Introduction

Online propaganda is potentially harmful to society, and the task of automated propaganda detection has been suggested to alleviate its risks (Martino et al., 2020b). In particular, providing a justification when performing propaganda detection is important for acceptability and application of the decisions. Previous challenges have focused on the detection of propaganda techniques (Martino et al., 2020a), based on news articles. However, many use cases do not solely involve text, but can also involve other modalities, notably images. Task 6 of SemEval-2021 proposes a shared task on the detection of persuasion techniques detection in memes, where both images and text are involved. Substasks 1 and 2 deal with text in isolation, but we focus on subtask 3: visuolinguistic persuasion technique detection.

This article presents the system behind our submission for subtask 3 (Dimitrov et al., 2021). To handle this problem, we use a model containing three components: data augmentation, image and text feature extraction, and chain classifier components. First, given a paired image-text as the input, we paraphrase the text part using back-translation and pair it again with the corresponding image to enrich the data. Then, we extract visual and textual features using the CLIP (Radford et al., 2021) image encoder and text encoder, respectively. Finally, we use a chain classifier to model the relation between labels for the final prediction. Our proposed method, named LIIR, has achieved a competitive performance with the best performing methods in the competition. Also, empirical results show that the augmentation approach is effective in improving the results.

The rest of the article is organized as follows. The next section reviews related works. Section 3 describes the methodology of our proposed method. We will discuss experiments and evaluation results in Sections 4 and 5, respectively. Finally, the last section contains the conclusion of our work.

2 Related work

This work is related to computational techniques for automated propaganda detection (Martino et al., 2020b) and is the continuation of a previous shared task (Martino et al., 2020a).

Taks 11 of SemEval-2020 proposes a more fine-graind analysis by also identifying the underlying techniques behind propaganda in news text, with annotations derived from previously proposed propaganda techniques typologies (Miller, 1939; Robinson, 2019).

This current iteration of the task tackles a more challenging domain, by including multimodal content, notably memes. The subtle interaction between text and image is an open challenge for state of the art multimodal models. For instance, the Hateful Memes challenge (Kiela et al., 2020) was recently proposed, as a binary task for detection of hateful content. The recent advances in pretraining of visuolinguistic representations (Chen et al., 2020) lead the model closer to human accuracy (Sandulescu, 2020).

More generally, propaganda detection is at the crossroad of many tasks, since it can be helped by many subtasks. Fact-checking (Aho and Ullman, 1972; Dale, 2017) can be involved with propaganda detection, alongside various social, emotional and discursive aspects (Sileo et al., 2019a),

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1015–1019
Bangkok, Thailand (online), August 5–6, 2021.

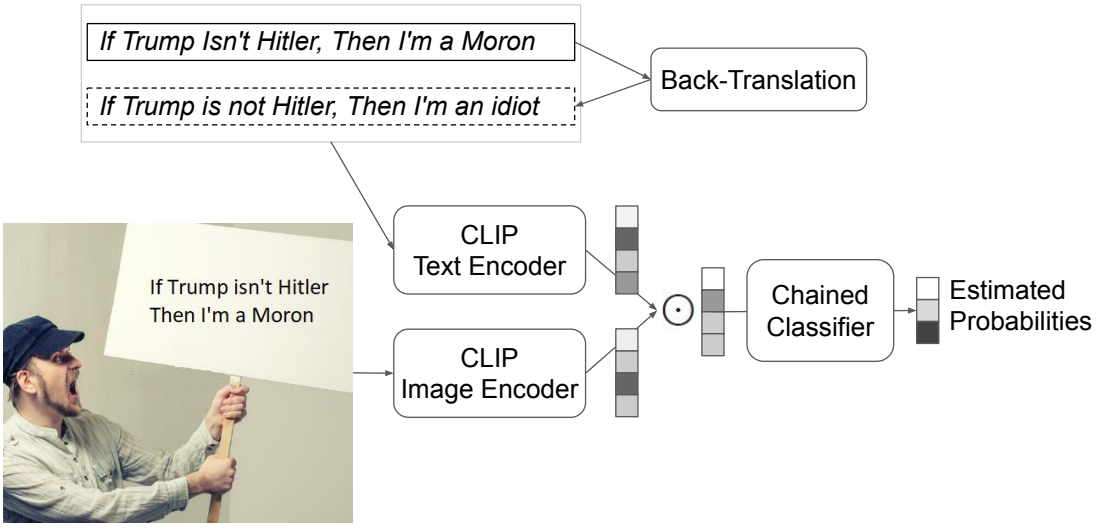

Figure 1: The overall architecture of our proposed model. For each example, use Back-Translation to derive augmentations of the text, and we compute persuasion techniques probabilities separately. Then, we average the estimated probabilities from augmented and original examples.

including offensive language detection (Pradhan et al., 2020; Ghadery and Moens, 2020) emotion analysis (Dolan, 2002), computational study of persuasiveness (Guerini et al., 2008; Carlile et al., 2018) and argumentation (Palau and Moens, 2009; Habernal and Gurevych, 2016).

3 Methodology

In this section, we introduce the design of our proposed method. The overall architecture of our method is depicted in figure 1. Our model consists of several components: a data augmentation component (Back-translation), a feature extraction component(CLIP), and a chained classifier. Details of each component are described in the following subsections.

3.1 Augmentation Method

One of the challenges in this subtask is the low number of training data where the organizers have provided just 200 training samples. To enrich the training set we propose to use the back-translation technique (Sennrich et al., 2016) for paraphrasing a given sentence by translating it to a specific target language and translating back to the original language. To this end, we use four translation models, English-to-German, German-to-English, English-to-Russian, and Russian-to-English provided by (Ng et al., 2019). Therefore, for each training sentence, we obtain two paraphrased version of it. In the test time, we average the probability distributions over the original and paraphrased sentence-image pairs.

3.2 Feature Extraction

Our system isProbabilities of a combination of pretrained visuolinguistic and linguistic models.

We use CLIP (Radford et al., 2021) as a pretrained visuolinguistic model. CLIP provides an image encoder f_i and a text encoder f_t. They were pretrained on a prediction of matching image/text pairs. The training objective incentivizes high values of $f_i(I).f_t(T)$ if I and T are matching in the training corpus, and low values of they are not matching[1]. Instead of using a dot product, we create features with element-wise product $f_i(I) \odot f_t(T)$ of image and text encoding. This enables aspect-based representations of the matching between image and text. We experimented with other compositions (Sileo et al., 2019b) which did not lead to significant improvement.

We then use a classifier C on top of $f_i(I) \odot f_t(T)$ to predict the labels.

3.3 Chained Classifier

In this task, we are dealing with a multilabel classification problem, which means we need to predict a subset of labels for a given paired image-text sample as the input. We noticed that label

[1]They assign each image to the text associated to other images in the current batch to generate negative examples

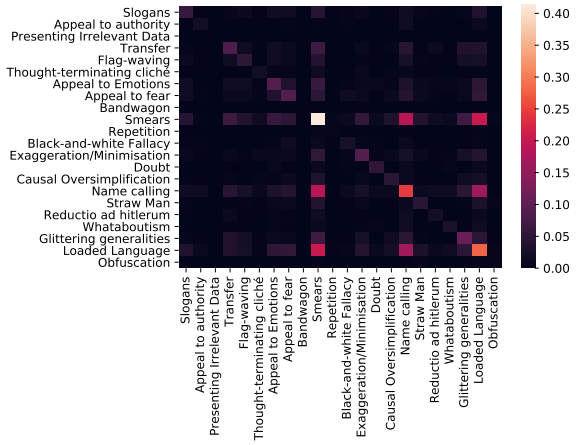

Figure 2: Probabilities of label co-occurence in the training set. Some label pairs, for instance (SMEARS and LOADED LANGUAGE) are frequently associated.

co-occurrences were not uniformly distributed, as shown in figure 2. To further address the data sparsity, we use another inductive bias at the classifier-level with a chained classifier (Read et al., 2009) using scikit-learn implementation (Pedregosa et al., 2011).

Instead of considering each classification task independently, a chained classifier begins with the training of one classifier for each of the L labels. But we also sequentially train L other classifier instances thereafter, each of them using the outputs of the previous classifier as input. This allows our model to model the correlations between labels. We use a Logistic Regression with default parameter as our base classifier.

Our chain classifier uses combined image and text features as the input. We transfer the predicted probabilities of the classifier via the sigmoid activation function to make the probability values more discriminating(Ghadery et al., 2018). Then we apply thresholding on the L labels probabilities since the task requires a discrete set of labels as output. We predict a label when the associated probability is above a given threshold. We optimize the threshold on the validation set by a simple grid search using values between 0.0 and 0.9 with a step of 0.005.

4 Experiments

4.1 Datasets

We use the dataset provided by SemEval-2021 organizers for task 6. The dataset consists of 687(290) samples as the training set, 63 samples as the dev set, and 200 samples as the test set. Each sample is an image and its corresponding text. We use 10% of the training set as the validation set for hyperparameter tuning.

Label	Count
Smears	199
Loaded Language	134
Name calling/Labeling	118
Glittering generalities (Virtue)	54
Appeal to (Strong) Emotions	43
Appeal to fear/prejudice	42
Exaggeration/Minimisation	42
Transfer	41
Slogans	28
Doubt	25
Flag-waving	24
Causal Oversimplification	22
Misrepresentation of Someone's Position	21
Whataboutism	14
Black-and-white Fallacy/Dictatorship	13
Thought-terminating cliché	10
Reductio ad hitlerum	10
Appeal to authority	10
Repetition	3
Obfuscation, Intentional vagueness, Confusion	3
Bandwagon	1
Presenting Irrelevant Data (Red Herring)	1

Table 1: Labels of persuasion techniques with associated counts in the training set

5 Evaluation and Results

5.1 Results

In this section, we present the results obtained by our model on the test sets for Subtask 3. Table 2 shows the results obtained by the submitted final model on the test set. All the results are provided in terms of macro-F1 and Micro-F1. Furthermore, we provide the results obtained by the random baseline, the best performing method in the competition, and median result for the sake of comparison. Note that, we used the first released training set at the time of final submission which contained just 290 training samples. Therefore, we also provide results obtained by our model after using all the provided 687 training samples. Results show that LIIR has achieved a good performance compared to the majority class baseline and the median result which demonstrates that our model can effectively identify persuasion techniques in text and images. Also, we can observe LIIR has achieved a competitive performance compared to the best result obtained by the best team in the competition when it uses all the training samples.

System	Macro-F1	Micro-F1
Majority class	0.05152	0.07062
Median	0.18842	0.4896
LIIR(290 examples)	0.18807	0.49835
LIIR(687 examples)	0.21796	0.51122
Best system	**0.27315**	**0.58109**

Table 2: The results obtained by LIIR compared to the baselines on the Test set for Subtask 3. Numbers in parentheses show the total number of train samples used by our model.

5.2 Ablation Analysis

In this part, we provide an ablation study on the effect of different components of our proposed method on the dev set. First, we show the effect of using just visual features, just textual features, and both. Furthermore, we examine how well the final results of our model was influenced by the augmentation method. Table 3 shows the ablation study on the effect of using different features. The first observation is that image features contain more information compared to the textual features. Also, we can observe that the best Micro-F1 score is obtained when we combine both visual and textual features. These results show the effectiveness of our method in making use of both visual and textual information.

System	Macro-F1	Micro-F1
LIIR – textual features	0.32275	0.53237
LIIR – visual features	**0.33347**	0.52954
LIIR	0.29972	**0.58312**

Table 3: Ablation analysis for the effect of using different features by our model on the dev set.

In Table 4, the effect of the augmentation technique is shown. As the results show, the augmentation approach is quite effective in improving the model performance by a high margin.

System	Macro-F1	Micro-F1
LIIR w/o Augmentation	0.25090	0.54952
LIIR w Augmentation	**0.29972**	**0.58312**

Table 4: Ablation analysis for the effect of augmentation method on the dev set.

6 Negative Results

We also tried to use CLIP as a zero-shot classifier for propaganda technique detection. To do so, we constructed prompts such as :

(1) *This image is committing* [LABEL] *fallacy.*

or

(2) *Saying that* [TEXT] *is* [LABEL] *fallacy.*

For each input image/text, we generated a prompt for each labels, and used CLIP to estimate the estimate an affinity score between the prompt and the image. CLIP is designed to predict relatedness between the input image and text, and we expected that an input text mentioning the relevant propaganda technique should be associated with higher probabilities that the others.

However, this method did not seem to perform better than chance. This suggests that propaganda detection technique task might be too abstract for CLIP in zero-shot settings.

7 Conclusion

We described our submission for the shared task of multimodal propaganda technique detection at SemEval-2021. Our system performances that are competitive with other systems even though we used a simple architecture with no ensemble, by leveraging non-supervised learning. We believe that further work on zero-shot learning would be a valuable way to improve propaganda detection techniques for the least frequent labels.

8 Acknowledgments

This research was funded by the CELSA project from the KU Leuven with grant number CELSA/19/018.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Robert Dale. 2017. NLP in a post-truth world. *Natural Language Engineering*, 23(2):319–324.

Dimiter Dimitrov, Giovanni Da San Martino, Hamed Firooz, Fabrizio Silvestri, Preslav Nakov, Shaden Shaar, Firoj Alam, and Bishr Bin Ali. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the Fifteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics.

Raymond J Dolan. 2002. Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194.

Erfan Ghadery and Marie-Francine Moens. 2020. Liir at semeval-2020 task 12: A cross-lingual augmentation approach for multilingual offensive language identification. *arXiv preprint arXiv:2005.03695*.

Erfan Ghadery, Sajad Movahedi, Heshaam Faili, and Azadeh Shakery. 2018. An unsupervised approach for aspect category detection using soft cosine similarity measure. *arXiv preprint arXiv:1812.03361*.

Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Resources for persuasion. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

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

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization. Survey track.

C.R. Miller. 1939. *How to Detect and Analyze Propaganda ...: An Address Delivered at Town Hall, Monday, February 20, 1939*. A Town Hall pamphlet. Town Hall, Incorporated.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. pages 314–319.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. In *Advances in Data and Information Sciences*, pages 433–439. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer.

Robert C Robinson. 2019. A rulebook for arguments, by a. weston. *Teaching Philosophy*, 42(4):425–428.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *CoRR*, abs/2012.13235.

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.

Damien Sileo, Tim Van de Cruys andmain Camille Pradel, and Philippe Muller. 2019a. Discourse-based evaluation of language understanding. *CoRR*, abs/1907.08672.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019b. Composition of sentence embeddings: Lessons from statistical relational learning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 33–43, Minneapolis, Minnesota. Association for Computational Linguistics.