Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification

Rajaraman Kanagasabai[1], Saravanan Rajamanickam[1], Hariram Veeramani[2], Adam Westerski[1], and Kim Jung Jae[1]
[1]Agency for Science, Technology and Research (A*STAR)
Institute for Infocomm Research, Singapore,
[2]University of California Los Angeles (UCLA), USA
[1]{kanagasa,saravananr,adam-westerski,jjkim}@i2r.a-star.edu.sg
[2]hariram@ucla.edu

Abstract

In this paper, we describe our system for ImageArg-2023 Shared Task that aims to identify an image's stance towards a tweet and determine its persuasiveness score concerning a specific topic. In particular, the Shared Task proposes two subtasks viz. subtask (A) Multimodal Argument Stance (AS) Classification, and subtask (B) Multimodal Image Persuasiveness (IP) Classification, using a dataset composed of tweets (images and text) from controversial topics, namely gun control and abortion. For subtask A, we employ multiple transformer models using a text based approach to classify the argumentative stance of the tweet. For sub task B we adopted text based as well as multimodal learning methods to classify image persuasiveness of the tweet. Surprisingly, the text-based approach of the tweet overall performed better than the multimodal approaches considered. In summary, our best system achieved a F1 score of 0.85 for sub task (A) and 0.50 for subtask (B), and ranked 2nd in subtask (A) and 4th in subtask (B), among all teams submissions.

1 Introduction

Persuasiveness mining is an important task within Argument Mining(Green, 2014; Stede et al., 2019), that is aimed at detecting and analyzing the ability to influence one's beliefs, attitude, intentions, motivation, and behavior (Lawrence and Reed, 2019). It has gained increased attention recently (Carlile et al., 2018; Chakrabarty et al., 2020) though most of the research focused on texts.

Persuasion, however, may depend not only on natural language but on other modalities (eg. visual means) as well. ImageArg is an initiative that attempts to capture this opportunity and expand persuasiveness mining into a multi-modal realm (Liu et al., 2022, 2023). It presents a multi-modal dataset consisting of annotations on tweets along with associated images, that supports benchmarking of state-of-the-art models on multiple argumentative classification tasks. ImageArg Shared Task 2023 proposes two subtasks viz. subtask (A) Multimodal Argument Stance (AS) Classification: Given a tweet composed of a text and image, predict whether the given tweet supports or opposes the topic, and subtask (B) Multimodal Image Persuasiveness (IP) Classification: given a tweet composed of text and image, predict whether the image makes the tweet more persuasive or not. In this paper, we report our systems for addressing both the subtasks.

Transformer based Multimodal text-embedded classification has been a promising approach recently (Sun et al., 2021; Liang et al., 2022b; Li et al., 2019; Radford et al., 2021; Li et al., 2019; Jia et al., 2021; Dosovitskiy et al., 2020). Taking inspiration from this, we explore multiple transformer models using text as well as multimodal learning methods, for both subtasks (A) and (B). Surprisingly, the text-based approach of the tweet performed better than the multimodal approaches considered. In particular, our best text based model achieved a F1 score of 0.85 for sub task (A) and 0.50 for subtask (B), and ranked 2nd in subtask (A) and 4th in subtask (B), among all teams submissions. Also, our benchmark results highlight the challenge of these tasks and indicate there is ample of room for model improvement. We demonstrate the limitation of these general multi-modal methods and discuss possible future work.

2 Related works

2.1 Stance Classification:

Stance Detection has been extensively studied in the literature ranging from detecting the stance of authors towards a single topic or different aspects of heterogeneous topics/entities (Küçük and Can, 2020). Some of the earlier contributions (Augenstein et al., 2016; Riedel et al., 2017; Thorne

Proceedings of the 10th Workshop on Argument Mining, pages 181–186

et al., 2017) to stance detection involved the usage of basic ML algorithms, bag-of-words(BOW) as features, TF-IDF feature based dense MLPs, sequence models such as LSTM by processing temporal and linguistic sequence information. Recently, several approaches have emerged adopting transformer based architectures. While stance detection is being actively pursued (Liang et al., 2022a), challenges such as the following remain: i) Learning with less data ii) Learning contrastive representations robust enough for complex stance features jointly by reusing the encoder representations to directly classify the stance based on extracted features as opposed to using a dedicated classifier, iii) Identifying right modality combination for the anchor, reference subspaces.

2.2 Persuasiveness Classification:

Past works have addressed several persuasiveness related tasks (Carlile et al., 2018; Chakrabarty et al., 2020), and in particular, ranking debate arguments (Wei et al., 2016), how audience variables (e.g., personality) influence persuasiveness through different argument styles (Lukin et al., 2017; Persing and Ng, 2017), but mainly focused on texts. (Nojavanasghari et al., 2016) explored coarse-grained fusion ideas such as concatenation for persuasiveness mining. In the area of vision-language, tasks are mainly designed for evaluating models' ability to understand visual information as well as expressing the reasoning in language (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019). In addition to the main stream, a few works study the relationship between image and text: (Alikhani et al., 2019) annotates the discourse relations between text and accompanying imagery in recipe instructions; and (Kruk et al., 2019) investigates the multi-modal document intent in instagram posts. However, multimodal learning for AM has been under-explored due to a lack of multi-modal corpora.

3 Task and Dataset Description

ImageArg dataset is composed of tweets (images and text) from controversial topics, namely gun control and abortion. ImageArg shared task is divided into two subtasks.

Subtask A: *Argumentative Stance (AS) Classification* Given a tweet composed of a text and image, predict whether the given tweet Supports or Opposes the given topic, which is a binary classification task.

Subtask B: *Image Persuasiveness (IP) Classification* Given a tweet composed of text and image, predict whether the image makes the tweet text more Persuasive or Not, which is also a binary classification task.

For convenience, below we refer to the subtasks (A) and (B) simply as Tasks A and B.

4 Our approach

4.1 Task A - Stance Classification:

For Task A, as the training data is not large, we ventured to explore a predominantly text-based approach, with tweet text and tweet image contents extracted from OCR fed as the inputs to the system. Our idea was to build a model capable of learning their corresponding unified representations which could be sufficiently discriminative in the stance detection classifier space. We considered multiple candidate models that satisfy this criteria and evaluated them on the ImageArg dataset. For all our approaches, we randomly split the instances into 80/20 percent and performed 5-fold cross-validation on the validation(dev) set to select the best model.

Approach 1: (T5 NLI) We used pretrained T5(Text-to-Text Transformer) to fine tune the model for the given dataset and also adjusted the hyper-parameters based on the best performance. During T5 training, we set the number of beams as 50 and the number of returned sequences as 5.

Approach 2: (BERTweet-based model) Sentiment based classifier using BERTweet(Nguyen et al., 2020), a large-scale language model pretrained for English Tweets using RoBERTa model and cross-entropy loss with custom linear layers. The positive and negative labels of the classifier corresponds to support and oppose labels of stance classification task. We have used the pretrained BERTweet model and fine-tuned the model for its best performance.

Approach 3: (Contrastive BERT model) : We adopt a multi-task contrastive learning framework with a two step representation learning paradigm, similar to (Chen et al., 2022). Firstly, stance label prefixed textual sequences were fed as inputs to a transformer encoder as the target Input anchor. Second, the corresponding positive and negative reference input samples were fed as inputs to a shared BERT encoder in the parameter space. Then, the final hidden state classifier token [CLS] is used as the

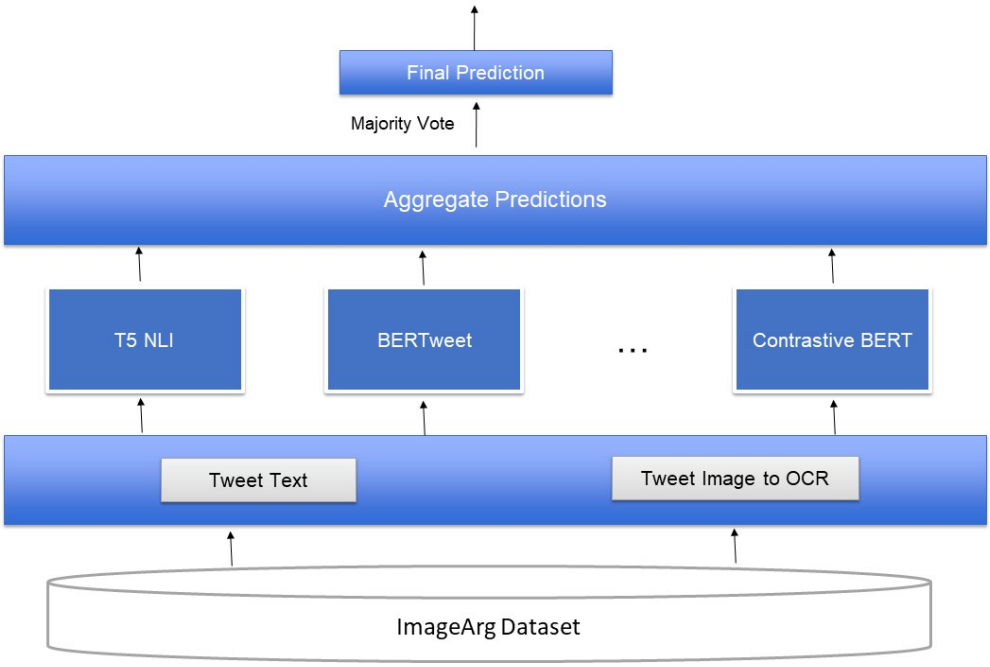

Figure 1: Illustration of Our Approach - Ensemble model of multiple classifiers such as T5 NLI, BERTWeet model, Contrastive BERT

standalone output label representation of the input sequence and the remaining hidden layer outputs served as sequence representation with labels encoded. Among the different variants of contrastive learning available, we used dual stance aware supervised contrastive learning technique with linear classifier. We also evaluate several candidate groups to serve as the anchor, positive, negative reference triplets in the proposed Dual Stance Aware Supervised Contrastive Learning space and found the more straightforward tweet text to act as efficient anchors in this space.

Approach 4: (Ensemble model) : We also considered a final model that uses an ensemble approach. In this model, we classify new data points by first applying the above 3 models, and then taking a majority vote of the predictions. In other words, the final prediction is determined by the class predicted by at least two models.

We have experimented a few other approaches, but as we observed on validation set, Contrastive BERT performed the best, followed by T5 and BERTweet based model. The ensemble model was marginally better in comparison. Thus we considered only these four models.

4.2 Task B - Image Persuasiveness:

For Task B, given the previously studied limitations in literature of projecting the claim and the evidence separately, it becomes imperative to utilize both the tweet text and the tweet image to assess the persuasiveness of the input sample. Hence, we propose separate models for Task B which can jointly deal with both the input modalities or the corresponding input sequences and understand their representations. Thus, as in Task A, we explored multiple candidate models and evaluated them on the ImageArg dataset.

Approach 1: (T5 NLI model) We adopted a sentence pair classification approach with T5 model. The tweet text and tweet image (OCR to Text) were passed as the two sentences, and fine tuned the model for the image persuasiveness dataset. We adjusted the hyper-parameters based on the best performance, as in the case of Task A.

Approach 2: (Stancy BERT) We use a BERT-base model which is fine-tuned with the standard Cross-Entropy Loss and the proposed consistency loss based on sequence similarity based on the tweet text evidence and supporting tweet image based texts/captions/expressions. This joint loss helps the model to acquire classifying features in addition to features central to stance similarity between two sequences.

Approach 3:(Multimodal ALBEF model) In addition to the text only model, we also experimented with multimodal fusion techniques using pretrained models of image encoder ResNet50 (He

et al., 2016), VGG and ViT, ALBEF model (aligning the image and text representations before fusing them through cross-modal attention with contrastive loss) and fine-tune them with linear classifiers. During validation, multimodal learning for Image Persuasiveness for Task B (Image and Text only) using ALBEF model performs better than the other variants of image encoders.

Approach 4: (Ensemble model) As in Task A, we considered an ensemble model that adopts a majority vote of the predictions by the 3 models above.

5 Experiments & Results:

For both Task A and Task B, we used tweet text, tweet image-text (OCR to text, using EasyOCR tool[1]) and custom pre-processing techniques to refine and clean up the textual sequences. For the latter, we used the BERTweet preprocessing scripts[2]. All the images are resized to (224*224) dimension and minor data augmentation(i.e., horizontal-flipped, rotation) was performed during training.Subsequently, we trained and performed experiments as outlined in Section 4.

To measure performance, we employ Precision, Recall and F1-score as metrics. In addition, we also consider class-weighted F1-score to account for class imbalance.

The experiments were executed on NVIDIA-GeForce Tesla V100 series SXM2-32GB with 5 cores of GPU machines. Models were trained for 10 epochs, and the pretrained weights for the transformers prior to fine-tuning were downloaded from the HuggingFace Library.

5.1 Task A - Stance Classification:

For stance classification, we adopted the four approaches described in Section 4.1. We used hyperparameters which were previously found to be optimum for Textual Entailment tasks including a Contrastive System Loss, AdamW optimizer, learning rate of 2e-5/5e-5.

The results of Task A on test set are shown in Tables 1 and 2.

Expectedly, the ensemble model achieved the best performance on the test also, but the ranking of the other models was slightly different. We argue that this could be because of the variance and the size of the dataset being on the smaller side.

Table 4 shows two examples where most of the models misclassified. In the first example, the summary text is inclusive but unfortunately requires the full text from the URL to classify correctly. The second example is expressed in a supportive tone, but the facepalm expression presumably misleads model to classify this as a sarcastic/opposing tweet.

5.2 Task B - Image Persuasiveness:

The Task B results on test are presented in Tables 1 and 3. We observe that, on test set, T5 NLI performed the best, followed by the ensemble model. The multimodal approach (ALBEF) had a surprisingly poor score, which could mean that larger datasets are required to deal with multimodal classification. Also, the results imply that Image Persuasiveness classification is a far more challenging problem and there is significant room for improvement.

6 Conclusion

This paper described our system for ImageArg-2023 Shared Task consisting of two subtasks viz. Subtask (A) Multimodal Argument Stance (AS) Classification, and Subtask (B) Multimodal Image Persuasiveness (IP) Classifi- cation. The tasks used a dataset composed of tweets (images and text) from controversial topics, namely gun control and abortion.

For subtask (A), we employ multiple transformer models using a text based approach to classify the argumentative stance of the tweet. For sub task (B) we adopted text based as well as multimodal learning methods to classify image persuasiveness of the tweet. Surprisingly, the text-based approach of the tweet overall performed better than the multimodal approaches considered. In summary, our best system achieved a F1 score of 0.85 for sub task (A) and 0.50 for subtask (B), and ranked 2nd in subtask (A) and 4th in subtask (B), among all teams submissions.

The results imply that image persuasiveness classification is a far more challenging problem and there is a significant room for improvement. However, it might require larger datasets to deal with the multimodal classification challenges.

[1]https://github.com/JaidedAI/EasyOCR

[2]https://github.com/VinAIResearch/BERTweet

References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A cor-

Task	Model name	F1 Positive(test)	F1 Weighted(test)
Stance Classification	T5 (NLI based)	0.8333	0.8533
	BERTweet+Linear layer+CE Loss	0.8429	0.8633
	BERT+Dual Contrastive Loss	0.8473	0.8604
	Simple Ensemble	**0.8504**	0.8504
Image Persuasiveness	T5 (NLI based)	**0.5022**	0.6300
	Stancy BERT	0.4123	0.5533
	Multimodal ALBEF model	0.2839	0.6466
	Simple Ensemble	0.4633	0.6833

Table 1: Task A: Stance Classification - F1 scores of submitted models and Task B: Image Persuasiveness - F1 scores of submitted models

Topic	F1 +ve	Precision	Recall
Abortion	0.7532	0.8788	0.6591
GunControl	0.8865	0.9647	0.8200

Table 2: Topicwise Results for Task A- Best Performing Model

Topic	F1 +ve	Precision	Recall
Abortion	0.4644	0.4603	0.4733
GunControl	0.5020	0.5233	0.5267

Table 3: Topicwise Results for Task B- Best Performing Model

Example	Incorrect Label
Poland 's anti-abortion push highlights pandemic risks to democracy HTTPURL HTTPURL. HTTPURL	Support
This packaging could be a useful form of gun control. Put all guns in these and no one will ever be able to get them out. **FACEPALM**. HTTPURL	Oppose

Table 4: Misclassified Examples- Sentences vs Incorrect Labels

pus of image-text discourse relations. *arXiv preprint arXiv:1904.06286*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

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.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2020. Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*.

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022. Dual contrastive learning: Text classification via label-aware data augmentation. *arXiv preprint arXiv:2201.08702*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Nancy Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Work-*

shop on Argumentation Mining, pages 11–18, Baltimore, Maryland. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. *arXiv preprint arXiv:1904.09073*.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.

Tao Liang, Guosheng Lin, Mingyang Wan, Tianrui Li, Guojun Ma, and Fengmao Lv. 2022b. Expanding large pre-trained unimodal models with multimodal information injection for image-text multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15492–15501.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. Imagearg: A multi-modal tweet dataset for image persuasiveness mining. *arXiv preprint arXiv:2209.06416*.

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *IJCAI*, pages 4082–4088.

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. In *International conference on machine learning*, pages 8748–8763. PMLR.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Manfred Stede, Jodi Schneider, and Graeme Hirst. 2019. *Argumentation mining*. Springer.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism*, pages 80–83.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.