IUST at ImageArg: The First Shared Task in Multimodal Argument Mining

Melika Nobakhtian, Ghazal Zamaninejad, Erfan Moosavi Monazzah, Sauleh Eetemadi
Iran University of Science and Technology
{melika.nobakhtian2000@gmail.com, gh_zamaninejad, moosavi_m@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

ImageArg is a shared task at the 10th ArgMining Workshop at EMNLP 2023. It leverages the ImageArg dataset to advance multimodal persuasiveness techniques. This challenge comprises two distinct subtasks: 1) Argumentative Stance (AS) Classification: Assessing whether a given tweet adopts an argumentative stance. 2) Image Persuasiveness (IP) Classification: Determining if the tweet image enhances the persuasive quality of the tweet. We conducted various experiments on both subtasks and ranked sixth out of the nine participating teams.

1 Introduction

Argumentation mining, a task in Natural Language Processing (NLP), aims to automatically detect argumentative structures in a document (Green et al., 2014). This process unveils not only people's viewpoints but also the reasons behind their beliefs (Lawrence and Reed, 2019). It offers valuable insights across a wide spectrum of fields, ranging from predicting financial market trends to public relations. However, prior research in this field mainly concentrates on text and does not exploit multimodal data.

ImageArg, a multimodal dataset introduced by Liu et al. (2022), is designed to bridge this gap. It includes persuasive tweets accompanied by images and its goal is to identify the image's stance towards the tweet and assess its persuasiveness score on specific topics.

ImageArg constitutes a collaborative challenge (Liu et al., 2023) tailored to advance multimodal persuasive techniques, using the ImageArg dataset. It is made of two subtasks: Argumentative Stance (AS) Classification and Image Persuasiveness (IP) Classification which will be further discussed in subsection 3.1 and subsection 3.2 respectively.

The whole system architecture is shown in Fig.1. We make three experiments on the AS subtask.

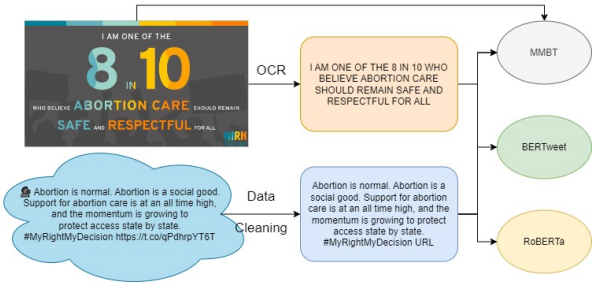

Figure 1: Our System Architecture

While in two of them, we only used the text as input data, in the third, we adopted a multimodal approach, considering both image and text inputs. In the former, we utilize BERTweet (Nguyen et al., 2020) in one experiment and RoBERTa (Liu et al., 2019) in the other. For our final experiment, we employed the Multimodal Bitransformer (MMBT) architecture (Kiela et al., 2020), harnessing tweets' text, text within images, and the images themselves. Our first approach, which leveraged BERTweet, achieved the highest F1-score compared to the two other methods.

For the IP task, we conducted a single experiment employing the MMBT model. We employed tweets' text, text extracted from images, and the images themselves as inputs.

2 Related Work

Persuasiveness Mining: Persuasiveness mining has been the subject of many recent studies (Chatterjee et al. (2014); Park et al. (2014); Lukin et al. (2017); Carlile et al. (2018); Chakrabarty et al. (2019)) but they do not provide the factors that make an argument persuasive. Liu et al. (2022) provides a framework to assign numerical score to the persuasiveness of an image based on its content type. They also determine the mode of persuasiveness for their images which can be based on reason, emotion, or ethics. In this work, we are going to use the dataset provided by Liu et al. (2022) for de-

Proceedings of the 10th Workshop on Argument Mining, pages 133–138

			Original			Processed		
	Topic	Split	Pos	Neg	Total	Pos	Neg	Total
Argumentative Stance	Gun Control	Train	475	448	923	470	442	912
		Dev	54	46	100	52	45	97
		Test	85	65	150	85	65	150
	Abortion	Train	244	647	891	**729**	644	**1373**
		Dev	19	81	100	19	81	100
		Test	33	117	150	33	117	150
Image Persuasiveness	Gun Control	Train	251	672	923	**747**	663	**1410**
		Dev	33	67	100	31	66	97
		Test	53	97	150	53	97	150
	Abortion	Train	278	613	891	**556**	609	**1165**
		Dev	26	74	100	26	74	100
		Test	53	97	150	53	97	150

Table 1: Statistics for the Original and Processed (Cleaning & Paraphrasing) Datasets. The 'Pos' class corresponds to 'Yes' and 'Support', while the 'Neg' class corresponds to 'No' and 'Oppose'. Numbers modified due to data augmentation are highlighted in bold.

termining image persuasiveness and argumentative mining.

Multimodal Learning: The recent surge in attention towards AI models lies in their capability to handle and comprehend inputs from multiple sources, thanks to the complementary nature of these multimodal signals in real-world applications (Aytar et al. (2016); Zhang et al. (2018); Alwassel et al. (2020)). Within the field of vision and language, tasks primarily revolve around assessing the models' proficiency in both grasping visual data and articulating reasoning through language (Agrawal et al. (2016); Goyal et al. (2017); (Hudson and Manning, 2019)). Although some research diverges from this mainstream which explores the connection between images and text: Alikhani et al. (2019) delve into annotating discourse relations between textual and accompanying visual elements in recipe instructions, while Kruk et al. (2019) delve into understanding multimodal document intent in Instagram posts.

3 Task and Data

ImageArg Shared Task includes two subtasks: Argumentative Stance (AS) Classification and Image Persuasiveness (IP) Classification. The dataset provided for this task encompasses two distinct topics of societal significance, namely abortion and gun control. Within the training subset of the dataset[1], a total of 912 examples are allocated to the domain of gun control, while 887 examples pertain to the topic of abortion. In the development subset, there are 100 data entries related to abortion and 97 data records related to gun control. In the testing partition, both the abortion and gun control categories are represented equally, each comprising 150 examples.

In the following parts, we will provide more details about subtasks and statistics related to the data specified for each subtask.

3.1 Argumentative Stance Classification

In this subtask, a tweet consisting of an image and text is given and the task is to predict whether this tweet supports or opposes a certain topic. It is considered a binary classification task; the proposed topics are abortion and gun control.

According to the data distribution shown in Fig.2 in the gun control section, we deal with a dataset that is approximately balanced and there is no need to worry about imbalanced classes. On the other hand, the abortion topic has different conditions; unfortunately, the dataset is imbalanced in both the train and dev sections. Over 70% of the data has been specified to the "Oppose" class.

3.2 Image Persuasiveness Classification

Like the previous subtask, a tweet composed of an image and text is given to a model as input and it will predict if the image beside the tweet text makes it more persuasive or not. The scenario is the same as the first subtask, a binary classification problem with the mentioned topics.

[1]We observed that we had data inconsistency according to the ImageArg statistics.

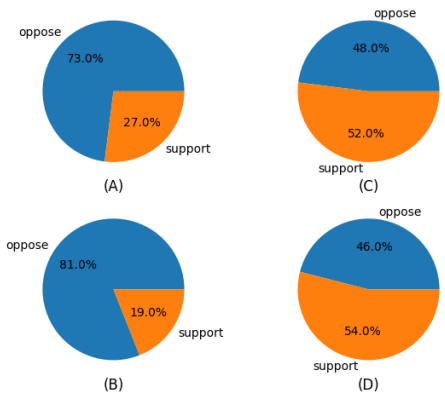

Figure 2: Data distribution in Argumentative Stance Classification. (A) Abortion Train. (B) Abortion Dev. (C) Gun Control Train. (D) Gun Control Dev.

As shown in Fig. 3, the dominant class label in both topics is "No", indicating that a significant portion of images does not enhance the persuasiveness of the tweet text. More than 65% of tweets on gun control and abortion belong to the "No" class.

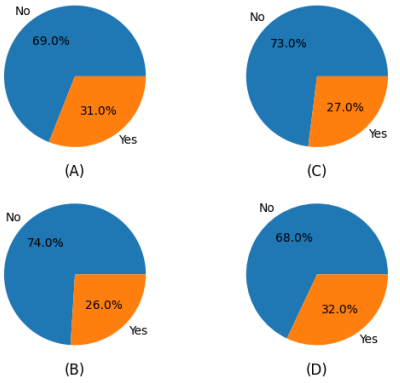

Figure 3: Data distribution in Image Persuasiveness Classification. (A) Abortion Train. (B) Abortion Dev. (C) Gun Control Train. (D) Gun Control Dev.

4 Methods

We first present preprocessing techniques used for both subtasks, as well as some ideas to make the performance of both tasks better before training models. Next, we introduce models developed for the argumentative stance followed by image persuasiveness models.

4.1 Preprocessing

Initially, we undertook text processing enhancements for the tweet content, incorporating various modifications to enhance their overall quality. In the preprocessed tweet corpus, all URLs were systematically substituted with the designated keyword "URL". A similar substitution approach was employed for mentions, seamlessly replaced by the keyword "MENTION". Given the inherent limitations of numerous text-processing models in deciphering emojis, a pragmatic approach of substituting them with the term "EMOJI" was adopted. Lastly, non-English characters were transcoded into their corresponding ASCII representations, subsequently utilized to supplant these characters within the text.

After inspecting the data instances, we found that many images have some text in their background. We assumed that including this text as an additional feature in the dataset, would improve our ability to develop more effective models for detecting valuable concepts. To achieve this, we used an OCR API[2] to extract text from images if it is available. It was the best tool that we came across in the variety of approaches.

OCR will bring many advantages to our approach. Firstly OCR can extract text from images that would otherwise be unavailable to the model. This can be especially useful for social media posts and other types of online content that often include images. Secondly, OCR can help to improve the performance of the model on multimodal data, where the image and the text are both relevant to the task.

In the preceding section, we examined the distribution of classes, revealing the presence of an imbalanced dataset issue. While diverse approaches exist to address this concern, our strategy is centered on employing oversampling techniques. Specifically, we chose to implement an oversampling methodology by augmenting the minority class instances independently for each subtask and topic. To achieve this equilibrium, we employed a paraphrasing technique facilitated by the ChatGPT paraphraser (Vladimir Vorobev, 2023) , harnessed from the foundational T5 model (Raffel et al., 2020). Tailored to each unique class ratio within varying contexts, a variable count of paraphrased samples was generated for each instance within the dataset. In the table 1 you can see the dataset statistics before and after applying pre-processing and paraphrasing techniques. Our primary objective was to approximate a balanced class distribution across diverse scenarios.

[2] https://ocr.space

4.2 Argumentative Stance Classification

We employed two different approaches for this subtask: One of them solely relies on text and the second method utilizes both images and text from tweets.

To ascertain the stance of tweets, it appears that placing trust in the textual content alone would suffice, given that images are unlikely to provide supplementary information. As a result, our first approach depends exclusively on text-based analysis. Within this approach, we used two distinct models for text classification, RoBERTa and BERTweet.

While RoBERTa leverages both the textual content of tweets and text extracted from accompanying images to infer stance, BERTweet focuses solely on training with tweet text. These models have undergone training on the entire dataset, encompassing gun control and abortion topics.

Our third approach capitalizes on a multimodal classification framework by integrating both textual content and images sourced from tweets. To realize this objective, we adopted the Multimodal Bitransformer (MMBT) architecture (Kiela et al., 2020), designed specifically to address image-and-text classification challenges. The MMBT model merges insights from text and image encoders. While the original configuration employs BERT (Devlin et al., 2019) as the text encoder and ResNet (He et al., 2015) as the image encoder, Inspired by (Neskorozhenyi, 2021) we replaced the image encoder with diverse iterations of the CLIP (Radford et al., 2021) model. CLIP, or Contrastive Language-Image Pre-Training, emerges as a neural network fine-tuned on (image, text) pairs, yielding feature representations that exhibit greater richness and applicability to the task at hand. Our exploration encompassed a spectrum of image encoders, loss functions, and optimizers within this framework, pursued to secure optimal outcomes for each distinct topic.

4.3 Image Persuasiveness Classification

Due to time limitations, we focused our efforts on presenting a singular methodology for this particular subtask. This approach harnesses the MMBT architecture, as detailed in the preceding section. This subtask similarly involves a multimodal classification challenge, entailing the utilization of both tweet images and text as inputs to the model. We undertook the development of separate models tailored to each individual topic, thereby enabling optimization specific to the nuances of each topic's content and characteristics.

Model	Topic	Precision	Recall	F1-score
BERTweet	All data	0.9068	0.6772	**0.7754**
	Abortion	0.8778	**0.5777**	**0.6824**
	Gun Control	0.9176	0.7358	0.8168
RoBERTa	All data	0.8475	**0.7143**	0.7752
	Abortion	0.8485	0.5600	0.6747
	Gun Control	0.8471	**0.8000**	**0.8229**
MMBT	All data	**0.9915**	0.3980	0.5680
	Abortion	**0.9697**	0.2222	0.3616
	Gun Control	**1.0000**	0.5667	0.7234

Table 2: Argumentative Stance classification results on test data

Model	Topic	Precision	Recall	F1-score
MMBT	All data	0.5000	0.4274	0.4609
	Abortion	0.5094	0.4030	0.4500
	Gun Control	0.4906	0.4561	0.4727

Table 3: Image Persuasiveness classification results on test data

5 Experiments and Results

First, we discuss our results of the first subtask, which is summarised in Tab.2. Our first and best submission for argumentative stance classification was BERTweet which is a variant of BERT specifically trained for tweets. We achieved 0.7754 F1-score on test data and we stand out as the 6-th team among others. BERT-based models are known for their strong performance in various NLP tasks, and this experiment confirms their utility for Argumentative Stance classification in tweets.

RoBERTa was the second submission and its result was highly close to BERTweet, with a score of 0.7752 based on F1. It suggests that incorporating text from images did not notably enhance the model's performance, which is an interesting finding. It is possible that the text within images may not have provided much additional useful information for this specific task. Both BERTweet and RoBERTa were trained for 10 epochs with batch-size of 8, using AdamW as optimizer (Loshchilov and Hutter, 2019).

MMBT was the last approach and it did not perform as well as the two first approaches. It yielded a noticeably lower F1-score of 0.5680 compared to the text-only models. Although we employed separate models for each topic, the image encoder was the same and we utilized CLIP-RN50x4 for this purpose. In addition, weighted Binary Cross

Entropy (BCE) was used as a loss function and we specified a weight according to class distribution for each topic for a better performance. The drop in performance could indicate that the addition of image information did not help and may have even introduced noise or complexity into the model. It's important to note that multimodal models can be challenging to train and may require a substantial amount of data and careful tuning to outperform text-only models in specific tasks.

Figure 4: Examples of Image Persuasiveness subtask that were misclassified

Figure 5: Examples of Image Persuassieveness that were classified correctly

In the second subtask, the only approach we followed was MMBT. The specified model for the gun control topic employed CLIP-RN101 as its image encoder whereas the abortion model used CLIP-RN50x16. These models were trained for 10 epochs with batch-size of 32. Its result is shown in Tab.3. While the model's performance may not be exceptionally high, it demonstrates some capability in assessing the persuasiveness of tweets with both text and image content. Some instances of the dataset with the model's predictions are shown in Fig.4 and Fig.5.

6 Conclusion

In this paper, we presented our approach in the ImageArg shared task which was the first shared task in Multimodal Argument Mining. We proposed three methods for the first subtask. These models have different varieties from models solely dependent on text to multimodal pre-trained models. We also had only one submission for the second subtask and we achieved 6-th place in both tasks among other groups that participated.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. 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 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video.

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.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Moitreya Chatterjee, Sunghyun Park, Han Suk Shim, Kenji Sagae, and Louis-Philippe Morency. 2014. Verbal behaviors and persuasiveness in online multimedia content. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media*

(SocialNLP), pages 50–58, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised multimodal bitransformers for classifying images and text.

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. 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 4622–4632, Hong Kong, China. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Rostyslav Neskorozhenyi. 2021. How to get high score using mmbt and clip in hateful memes competition. https://towardsdatascience.com/how-to-get-high-score-using-mmbt-and-clip-in-hateful-memes-competition-90bfa65cb117. Accessed: 2023-08-31.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *CoRR*, abs/2005.10200.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 50–57, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text.