GC-Hunter at ImageArg Shared Task: Multi-Modal Stance and Persuasiveness Learning

Mohammad Shokri
The Graduate Center
City University of New York
mmshokri@gradcenter.cuny.edu

Sarah Ita Levitan
Hunter College
City University of New York
sarah.levitan@hunter.cuny.edu

Abstract

With the rising prominence of social media, users frequently supplement their written content with images. This trend has brought about new challenges in automatic processing of social media messages. In order to fully understand the meaning of a post, it is necessary to capture the relationship between the image and the text. In this work we address the two main objectives of the ImageArg shared task. Firstly, we aim to determine the stance of a multi-modal tweet toward a particular issue. We propose a strong baseline, fine-tuning transformer based models on concatenation of tweet text and image text. The second goal is to predict the impact of an image on the persuasiveness of the text in a multi-modal tweet. To capture the persuasiveness of an image, we train vision and language models on the data and explore other sets of features merged with the model, to enhance prediction power. Ultimately, both of these goals contribute toward the broader aim of understanding multi-modal messages on social media and how images and texts relate to each other.

1 Introduction

Argumentative stance detection is an important problem within the field of natural language processing (NLP). Its primary objective is to discern the underlying position of a text in relation to a specific topic. Accurate identification of a text's stance enhances the performance of several other NLP applications, including text summarizing, information retrieval, fact-checking, and broadly contributes to enhanced understanding of the text. In recent years, the landscape of information dissemination has evolved beyond text, and a growing number of online users express themselves on social media using multi-modal messages. This shift underscores the need for more sophisticated approaches in argumentative stance detection .

The emergence of pre-training models based on transformer architecture (Vaswani et al., 2017) has introduced new horizons for analyzing and understanding text. All areas of natural language processing have been impacted by transformers and the subsequent models derived from them. Although remarkable strides have been made in most uni-modal applications of language processing, researchers are now shifting to multi-modal problems such as vision-language learning. Similar to the uni-modal challenges, large high quality labeled datasets are needed to pre-train the multi-modal models and fine-tune them for the downstream task.

In this work, we use lightweight vision and language learning models to learn joint representations of image-text pairs to capture patterns that help us predict how an image contributes to persuasiveness of a tweet comprised of an image and text. In addition to multi-modal models, we argue that to capture the stance of a tweet toward a given topic, only processing the text modality acts as a strong baseline for any multi-modal learning models. This is because detecting the stance of the text will provide valuable insights into the overall stance of the tweet itself.

The remainder of the paper is organized as follows. We first review previous studies on argument mining and stance detection, as well as vision and language learning in Section 2. Then we introduce the dataset used in this work in Section 3. In Section 4, we detail our experiments and results obtained. Finally, we conclude in Section 5 and summarize the main findings of this work.

2 Related Work

Numerous works have studied the problem of classifying argumentative stance, focusing on developing robust and accurate models for identifying the stance expressed in text. Existing studies have explored a different approaches, such as feature based classification, structure based classification, neural networks and attention based models, and domain specific knowledge and lexicons (Li and

Proceedings of the 10th Workshop on Argument Mining, pages 162–166

Caragea, 2019; Du et al., 2017; Rajadesingan and Liu, 2014; Habernal and Gurevych, 2017).

Earlier studies of argument mining mostly focused on learning the argumentative structure of a text document or classifying different argumentation strategies. Recently, researchers have begun to study persuasiveness-related tasks related to argument mining. Wei et al. (2016) proposed several features to capture persuasiveness in online forums. They argue that online persuasive texts contain an argument strategy that is not common in other genres. In a similar study, researchers created an annotated dataset comprised of argumentative text pairs on the same topics and performed a thorough analysis of how to quantify each argument's persuasiveness (Habernal and Gurevych, 2016). Despite these efforts to develop methods to identify persuasiveness of arguments in text, studying image persuasiveness is a largely unexplored problem.

Multi-modal learning involves joint processing of information from two or more modalities. In recent years, multi-modal learning has gained substantial attention in the machine learning community. Researchers have explored various architectures to effectively fuse information from different modalities. Some successful models use separate embeddings for image and text modalities and then capture the similarities using dot products or attention models (Faghri et al., 2017; Radford et al., 2021). Other models use deeper networks to model the image-text representations (Nguyen et al., 2020). In this work we build on prior studies to explore models and approaches for multi-modal stance detection. We use lightweight neural models for learning joint embeddings of image-text pairs in the data. We also capture similarity scores with more computationally expensive transformer embedders to gain more information about how both modalities interact with the given topic.

3 Dataset

We use the data provided for the ImageArg Shared Task 2023 (Liu et al., 2022, 2023). The data consists of a multi-modal corpus (ImageArg) of tweets on two social topics, *gun control* and *abortion*. The corpus was collected with the aim of studying the persuasiveness of a post that contains both text and an image, and also the argumentative stance of multi-modal tweets towards the topic. They develop schemes to annotate images based on their stance and persuasiveness. While stance detection is an established discipline with many resources, persuasiveness in a multimodal context is an underexplored problem without existing labeled corpora. To annotate the stance of the tweets, tweets are assumed to hold a consistent stance towards the topic in both modalities. The pipeline to annotate persuasiveness is designed in a way that only the tweets that have a clear stance towards the topic are annotated for how persuasive they are. The corpus is divided into train, validation, and test sets. The dataset details are provided in Table 1.

Topic	Train	Validaiton	Test
Abortion	891	100	150
Gun Control	923	100	150

Table 1: ImageArg dataset splits.

The train datasets are slightly imbalanced with regard to persuasiveness labels. Both datasets have more instances where the image is not making the tweet more persuasive. In terms of supporting or opposing the stance however, gun control dataset is quite balanced but in abortion dataset, "oppose" is the dominant class.

Dataset	Support	Oppose	Not Persuasive	Persuasive
Abortion	244	647	613	278
Gun Control	475	448	672	251

Table 2: Counts of labels in train datasets.

4 Experiments

4.1 Sub-task A

The aim of Subtask A of this shared task is to determine if a tweet composed of image and text supports or opposes a given topic, which is a binary classification problem. After carefully examining the data and the challenge, we hypothesized that a transformer based model fine-tuned only on text would be a solid baseline. That is because we expect users to express their attitude toward a topic in the written text and include pictures and graphics that further enhance their argument. We believe it's unlikely that a user would post an image that contradicts the message conveyed through the text. Therefore, we began our experiments by fine-tuning a BERT model (Devlin et al., 2018) on the tweet texts. We trained the model with a linear layer on top of it. We trained the model for ten epochs with a learning rate of $5e-5$, and saved

the best model at the end, based on their performance on the validation set. The results of the BERT model, evaluated on both abortion and gun control datasets, are shown in Table 3. As shown in the table, the model performs slightly better on gun control validation data than abortion data. This is possibly due to the more balanced nature of the gun control data. The abortion validation data mostly contains oppose labels. The model's F1 score on the combined test sets was 0.776.

Dataset	Class	Precision	Recall	F1	Support
Abortion	Support	0.92	0.63	0.75	19
	Oppose	0.92	0.99	0.95	81
	Macro Avg	0.92	0.81	0.85	100
	Weighted Avg	0.92	0.92	0.91	100
Gun Control	Support	0.89	0.90	0.91	52
	Oppose	0.90	0.86	0.88	44
	Macro Avg	0.90	0.89	0.89	96
	Weighted Avg	0.90	0.90	0.90	96

Table 3: Bert model fine-tuned on ImageArg validation data (Subtask A).

Next, we aimed to improve the results of our text-only classification. We fine-tuned an XLM-Roberta (Conneau et al., 2019) model on the data, as its pre-trained on a lot more data and its shown to outperform BERT on the GLUE benchmark (Wang et al., 2018). We trained the model for ten epochs with learning rate of $5e-6$, and saved the model with the best performance on the validation set. The scores are depicted in Table 4. This model boosted our scores significantly on both datasets. It also scored higher on the test set with an F1 score of 0.805.

Dataset	Class	Precision	Recall	F1	Support
Abortion	Support	1.00	0.63	0.77	19
	Oppose	0.92	1.00	0.96	81
	Macro Avg	0.96	0.82	0.87	100
	Weighted Avg	0.94	0.93	0.92	100
Gun Control	Support	0.96	0.88	0.92	52
	Oppose	0.88	0.95	0.91	44
	Macro Avg	0.92	0.92	0.92	96
	Weighted Avg	0.92	0.92	0.92	96

Table 4: XLM-Roberta model fine-tuned on ImageArg validation sets (Subtask A).

After training and evaluating our baseline models, we explored using other features which could capture possible helpful information in the data. We hypothesized that if a picture accompanies text in a tweet, it should have high similarity with some aspects of the topic. We gathered text-image similarity scores with VLP (Vision and Language Pre-training) models such as CLIP(Contrastive Language-Image Pre-Training (Radford et al., 2021)). Clip is a neural model developed by OpenAI and its innovation lies in its ability to learn meaningful associations between pairs of image and their corresponding textual description through a contrastive learning approach. However, combining these scores with the logits from our text-only transformer models did not seem to improve the results in neither topics. Our best results were obtained from a random forest classifier trained on the data using ViLT (Vision and Language Transformer) classification logits (Kim et al., 2021), CLIP similarity scores, and text similarity scores between tweet text and image text. The results are depicted in table 5. ViLT has a simple architecture for joining vision-language learning and has an efficient runtime due to its lightweight and convolution-free processing of pixel-level embeddings. Figure 1 shows how a Vilt model differs from other popular multi-modal models.

Dataset	Class	Precision	Recall	F1	Support
Abortion	Support	0.44	0.95	0.60	19
	Oppose	0.98	0.72	0.83	81
	Macro Avg	0.71	0.83	0.71	100
	Weighted Avg	0.88	0.76	0.79	100
Gun Control	Support	0.79	0.85	0.81	52
	Oppose	0.80	0.73	0.76	44
	Macro Avg	0.79	0.79	0.79	96
	Weighted Avg	0.79	0.79	0.79	96

Table 5: Best Random Forest model on trained with ViLT logits, CLIP scores, and text similarity scores (Subtask A).

4.2 Sub-task B

The goal of Subtask B is to predict whether an image makes the tweet text more persuasive or not. For instance, an image that is not related to the topic will not improve the persuasiveness of the tweet. In our initial analysis of the data, we observed that many pictures have some text written in them. Therefore, for our baseline submission to Subtask B, we began by using Python's EasyOCR[1] framework with the default recognition models to extract the texts in the images. We hypothesized that if the image contributed to the persuasiveness of the post, the image text should have high similarity scores to the tweet text. We then concatenated the image text with the tweet text, using a <SEP> token to separate them for the model input.

We trained a ViLT(Vision and Language Transformer) model on our data. We trained the model

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

separately for the two topics to maximize performance per topic. We experimented with training the Vilt on each dataset for 8 epochs and validating on the validation set. We used an Adam optimizer with a learning rate of $5e-5$. The results of our best model on the datasets are depicted in Table 6. It is clear from the results in the table that the model tends to learn better when an image does not make the tweet more persuasive. This is possibly due to the fact that it is the dominant class in the training set (Table 2).

Dataset	Class	Precision	Recall	F1	Support
Abortion	No	0.79	0.89	0.84	74
	Yes	0.50	0.31	0.38	26
	Macro Avg	0.64	0.60	0.61	100
	Weighted Avg	0.71	0.74	0.72	100
Gun Control	No	0.88	0.68	0.77	65
	Yes	0.54	0.81	0.65	31
	Macro Avg	0.71	0.74	0.71	96
	Weighted Avg	0.77	0.72	0.73	96

Table 6: Trained ViLT model performance on validation sets.

We also trained additional models with other features. For example, we ran a CLIP model on our data. The CLIP model only expects 77 tokens as text, which is the default value of the model and larger values are not supported by the model. Therefore, we passed the first 77 tokens of the text into the model and retrieved the similarity scores between the image and the tweet text. We then concatenated the image-text similarity scores with our previously extracted tweet text-image text similarity scores and passed them to a one-layered neural network along with the last hidden states of the ViLT model. We aimed to capture all the similarities among the text pairs and image-text pairs in the data. However, this model did not perform as well as our baseline on the validation set, so we did not submit it for the final evaluation.

We also trained another variation of the CLIP model on the data but passed only the context ("gun control" or "abortion") instead of the first 77 tokens of the tweet. We tested training another one-layered neural network with only the CLIP similarity scores and also merged it with the tweet text-image text similarity scores. Neither experiment outperformed our baseline scores on validation and test set. Our results on the test sets are shown in Table 7.

5 Conclusions

In this paper, we presented a strong model for multi-modal stance detection towards a given topic.

Topic	Precision	Recall	F1
Abortion	0.33	0.27	0.29
Gun Control	0.46	0.49	0.47

Table 7: Topic-wise results of our model on the test set.

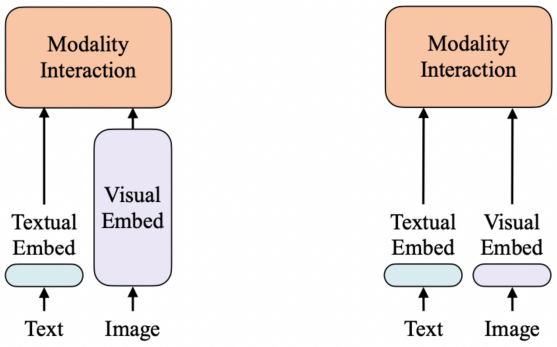

Figure 1: A ViLT model is shown on the right side, highlighting that it has fewer computations for extracting visual embeddings. It is compared with most vision language learning models that usually have an architecture more similar to the graph on the left. Figure taken from (Kim et al., 2021)

Our text-only fine-tuned models outperformed half of the participant teams, suggesting that that fine-tuning a transformer-based model only on tweet text could be a strong baseline for learning stance in multi-modal posts. To examine how an image contributes to persuasiveness of a tweet, we experimented with image-text similarity scores from a CLIP model, along with the similarity between any text in the image and the tweet text. We also extracted similarity scores between the image and the topic as another feature. Although this set of features did not produce the best results, future work could further explore these features and different ways of modeling them for improved performance.

Limitations

A limitation of our work, particularly for Subtask A, is that we did not fully explore multi-modal features. Because our text-only results outperformed our other experiments with image embeddings, we focused on those and did not explore further to extract helpful information from the image-text interaction. It is possible that a deeper exploration of both image and text modalities would yield better performance because it leverages the multimodal nature of the dataset.

Ethics Statement

This work has potential benefits that come along with potential risks. Social media platforms could benefit from a system that could perfectly detect the stance of posts towards sensitive topics that may affect the community's safety and well being, and possibly warn users or take action aligned with the guidelines of the platform. However, a system's failure to accurately identify stances or persuasive intent could inadvertently suppress genuine discourse by flagging legitimate viewpoints as misleading or manipulative, thus undermining freedom of expression. It is important for such models and systems to be interpretable and explainable so that decisions are not made based on black box systems.

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. *CoRR*, abs/1911.02116.

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*.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. 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 6299–6305.

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*.

Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. 2020. Movie: Revisiting modulated convolutions for visual counting and beyond. *arXiv preprint arXiv:2004.11883*.

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.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in twitter debates. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings 7*, pages 153–160. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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.