

Overview of ImageArg-2023: The First Shared Task in Multimodal Argument Mining

Zhexiong Liu, Mohamed Elaraby*, Yang Zhong*, Diane Litman

Department of Computer Science

University of Pittsburgh, Pittsburgh, PA 15260 USA

{zhexiong.liu, mse30, yaz118, dlitman}@pitt.edu

Abstract

This paper presents an overview of the *ImageArg* shared task, the first multimodal Argument Mining shared task co-located with the 10th Workshop on Argument Mining at EMNLP 2023. The shared task comprises two classification subtasks - (1) Subtask-A: Argument Stance Classification; (2) Subtask-B: Image Persuasiveness Classification. The former determines the stance of a tweet containing an image and a piece of text toward a controversial topic (e.g., gun control and abortion). The latter determines whether the image makes the tweet text more persuasive. The shared task received 31 submissions for Subtask-A and 21 submissions for Subtask-B from 9 different teams across 6 countries. The top submission in Subtask-A achieved an F1-score of 0.8647 while the best submission in Subtask-B achieved an F1-score of 0.5561.

1 Introduction

Research in Argument Mining (AM) typically centers around the examination of an author’s argumentative position, achieved through the automated identification of argument structures. This research has predominantly concentrated on domains presented in textual formats, encompassing endeavors such as mining persuasiveness in essays (Stab and Gurevych, 2014) and user-generated web discourse (Habernal and Gurevych, 2017). Recently, there has been a growing recognition of the need for multimodality in AM research. A noteworthy development in this regard is the *Retrieval for Argument* shared task (Carnot et al., 2023). This task is designed to retrieve images related to a controversial topic that aligns with the textual stance, whether it supports or contradicts the topic. In a related context, Liu et al. (2022) introduced the *ImageArg* corpus, which is designed to investigate multimodal persuasiveness within tweets. This corpus represented an advancement in the field of automated

* These authors contributed equally to this work.

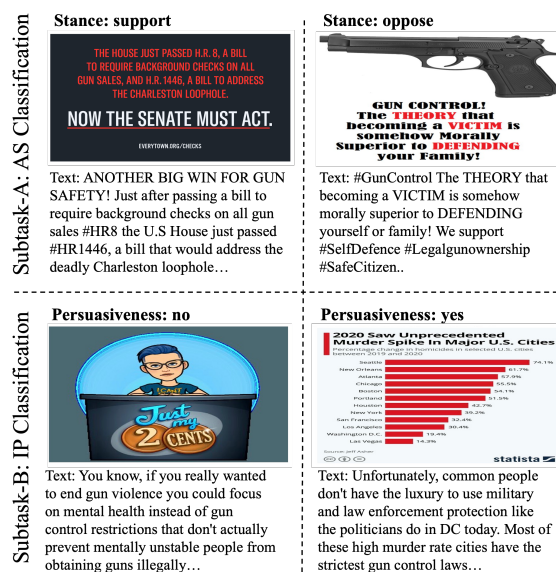


Figure 1: Examples of Subtask-A: Argument Stance (AS) Classification and Subtask-B: Image Persuasiveness (IP) Classification.

persuasive text identification (Duthie et al., 2016) by introducing a new modality through the inclusion of images.

This paper introduces the *ImageArg* shared task¹, building upon the groundwork laid by Liu et al. (2022) and conducted as a part of the 10th Workshop on Argument Mining². The shared task comprises two subtasks that center around two highly controversial topics (gun control and abortion):

- Subtask-A: Argument Stance (AS) Classification. The primary objective is to determine, for each of these topics, whether a given tweet text and its accompanying image express either support or opposition. This subtask addresses the research question: how to identify an argument stance of the tweet that contains a piece of text and an image?

¹<https://imagearg.github.io/>

²<https://argmining-org.github.io/2023/>

- **Subtask-B: Image Persuasiveness (IP) Classification.** The goal is to assess whether the image associated with a tweet makes the tweet text more persuasive or not. This subtask addresses the research question: does the tweet image make the tweet text more persuasive?

Figure 1 shows examples of the two subtasks. The upper left tweet expresses a strong stance towards supporting gun control by indicating a house bill about the requirement of background checks for all gun sales. The upper right tweet opposes gun control because it is inclined to self-defense. The lower left tweet has an image irrelevant to the gun control topic. It does not improve the persuasiveness of the tweet text (and its stance) that argues to focus on mental health instead of gun restriction. The lower right tweet makes the tweet text (and its stance) more persuasive because it provides strong evidence to show the statistics of the murder rate in major U.S. cities due to restrictive gun control laws, so citizens cannot easily arm themselves.

The shared task received 31 submissions for Subtask-A and 21 submissions for Subtask-B from 9 diverse teams, comprising both academic experts from various universities and industry researchers, across 6 different countries. In general, the submissions that utilized text information from tweet images and performed data augmentation yielded favorable results for Subtask-A. The submissions that utilized unified multimodal models also achieved good performance in Subtask-B. The highest Subtask-A F1-score of 0.8647 was attained by **Team KnowComp** (Zong et al., 2023), while the leading Subtask-B F1-score of 0.5561 was attained by **Team feeds** (Torky et al., 2023). Details about task submissions are described in Section 4.

2 Related Work

Multimodal Learning Recently, there has been increasing attention to assessing the ability of artificial intelligence models to process and understand multimodal input signals that occur in real-world applications (Zhang et al., 2018; Alwassel et al., 2020). In the vision-language domain, tasks are primarily designed to evaluate the capacity of models to comprehend visual data and articulate reasoning in language (Goyal et al., 2017; Hudson and Manning, 2019). In addition, Zheng et al. (2021) are interested in the discourse relations between text and its associated images in recipes, while Kruk et al. (2019) explores the multimodal document intent of

Instagram posts. More recently, Liu et al. (2022) introduce *ImageArg*, the first multimodal learning corpus for argument mining. However, the size of the *ImageArg* corpus is small, which motivates our construction of an extension of the original corpus. Regarding multimodal modeling, researchers have developed methods to derive strong representations for each modality and implement fusion techniques (Tsai et al., 2018; Hu et al., 2019; Tan and Bansal, 2019; Lu et al., 2020). Although several shared tasks in machine translation (Specia et al., 2016; Barrault et al., 2018) and argument retrieval (Carnot et al., 2023) have revealed the effectiveness of multimodal learning, none of them focused on argument persuasiveness. Therefore, this shared task provides opportunities to benchmark the new multimodal argument persuasiveness corpus by utilizing various image and text encoders along with effective fusion strategies.

Computational Persuasiveness While classical argument mining primarily focuses on the identification of argumentative components and their corresponding relationships (Stab et al., 2014, 2018; Lawrence and Reed, 2020), researchers have also focused on argument persuasiveness (Chatterjee et al., 2014; Park et al., 2014; Lukin et al., 2017; Carlile et al., 2018; Chakrabarty et al., 2019). Furthermore, while Riley (1954), O’Keefe (2015), and Wei et al. (2016) investigated the ranking of debate arguments on the same topic, they did not focus on discovering factors contributing to the persuasiveness of these arguments. In addition, Lukin et al. (2017) and Persing and Ng (2017) investigate how audience personality influences persuasiveness through diverse argument styles, such as factual versus emotional arguments. However, their work only focuses on the textual modality. In contrast, Higgins and Walker (2012) and Carlile et al. (2018) focus their attention on persuasion strategies, e.g., Ethos (credibility), Logos (reason), and Pathos (emotion), within the context of reports and student essays. Building upon their work designed for textual corpora, Liu et al. (2022) extend the annotation schemes to include the image modality. Although Park et al. (2014), Joo et al. (2014), and Huang and Kovashka (2016) employ facial expressions and bodily gestures to analyze persuasiveness within the realm of social multimedia, their investigations remain limited to human portraits and fail to generalize across diverse image domains. While prior work does explore persuasive advertisements

Confidence	Abortion	Gun control
>= L5	0.8437	0.7434
>= L4	0.7842	0.6697
>= L3	0.7824	0.6551
>= L2	0.7820	0.6516
>= L1	0.7807	0.6487

Table 1: Krippendorff’s alpha for abortion and gun control topics with respect to different confidence levels.

in a multimodal fashion (Hussain et al., 2017; Guo et al., 2021), it is important to note that their focus is on sentiment analysis, intent reasoning, and persuasive strategies tailored specifically for advertisements. In contrast, our shared task is interested in argument mining, marking an aligned goal to the *ImageArg* work (Liu et al., 2022), offering substantial value to multimodal computational social science.

3 Corpus

We extended the *ImageArg* corpus (Liu et al., 2022) by following its annotation protocol to annotate new data on abortion and gun control topics. Specifically, we annotated 1141 new abortion tweets and 301 new gun control tweets. Parts of the new gun control tweets were used to replace 131 out of the original 1003 gun control tweets in the *ImageArg* corpus which were no longer available due to deletions or account suspensions. The other extras were annotated to ensure gun control and abortion tweets have close data distributions. Therefore, we obtained 1173 gun control tweets in total. In addition to using the original annotation protocol (Liu et al., 2022), we required annotators to score confidence levels, which was designed to improve the inter-annotation agreement. Confidence was divided into 5 levels: L5-Extremely confident (understood and answered all annotations carefully), L4-Quite confident (tried to understand and answered most annotations carefully), L3-Somewhat confident (confused about some annotations), L2-Not very confident (did not understand some annotations), and L1-Not confident (mostly educated guesses).

In the annotation process, each tweet was annotated by three annotators on Amazon Mechanical Turk (AMT)³ who had done more than 5,000 approved annotations with at least 95% approved rates in their historical hits. Annotators were required to pass a qualification exam that annotated

³<https://www.mturk.com/>

Topic	Split	AS		IP		Total
		Sup.	Opp.	Yes	No	
Gun control	train	475	448	251	672	923
	dev	54	46	33	67	100
	test	85	65	53	97	150
Abortion	train	244	647	278	613	891
	dev	19	81	26	74	100
	test	33	117	53	97	150

Table 2: The data statistics for Subtask-A and Subtask-B for gun control and abortion topics.

pilot examples with at least 0.7 accuracy. Table 1 shows AS annotation agreements in terms of Krippendorff’s alpha (Krippendorff, 2011) and confidence levels. We observed that annotations with high confidence levels had high agreements but dropped more annotations. To make the trade-off between annotation costs and agreements, we disregarded annotations with confidence levels less than L4 for abortion and less than L5 for gun control. The remaining new AS annotations for abortion and gun control have alpha scores of 0.78 and 0.74, respectively. The new IP annotations were also inherited from the *ImageArg* protocol. First, annotators annotated two persuasiveness scores: one for tweet text (s_t), another for tweet text and image (s_{it}). Then we computed a score difference $\Delta s_i = \max(s_{it} - s_t, 0)$ as a persuasiveness gain from adding a tweet image. The final image persuasiveness score for each tweet was the average of persuasiveness gains from three annotators. To interpret image persuasiveness, we used the same threshold (0.5) in *ImageArg* to split them into binary labels, indicating whether the image made the tweet text more persuasive or not.

We split the corpus into train, development, and test sets in the shared task, which obtained 1814 train, 200 development, and 300 test samples for both subtasks⁴. The data statistics are shown in Table 2 for Subtask-A and Subtask-B, respectively. We released the train and development data splits for model development and the test set without labels before the task submission deadline. We shared the complete test set with labels after completing the shared task. The full corpus can be downloaded from the GitHub repository⁵.

⁴We removed one abortion tweet in the test set when we evaluated team submissions for the leaderboard because the tweet was no longer available during the task submission phase so a few teams were unable to download the full 300 test samples.

⁵<https://github.com/ImageArg/ImageArg-Shared-Task>

ID	System	Score	Modality	Model	Notes
1*	KnowComp-4	0.8647	I+T	ResNet50 + DeBERTa	Augment Text with Back Translation + WordNet
2	KnowComp-5	0.8571	I+T	ResNet50 + DeBERTa	Augment Text with Translation + WordNet + Semantic SimilarityAttention
3	KnowComp-1	0.8528	I+T	ResNet101 + DeBERTa	Augment Text with Translation + WordNet
4*	Semantists-4	0.8506	T+E		Ensemble of All Models
5	Semantists-3	0.8462	T+E	BERTweet	OCR on Image
6	Semantists-5	0.8417	T+E	BERT	Dual Contrastive Loss + OCR on Image
7	Semantists-1	0.8365	T+E	BERT	Contrastive Loss + OCR on Image
8	Semantists-2	0.8365	T+E	T5	OCR on Image
9	KnowComp-2	0.8365	I+T	ResNet50 + DeBERTa	Augment Text with Translation + WordNet + Semantic SimilarityAttention
10	KnowComp-3	0.8346	I+T	LayoutLMv3 + DeBERTa	Augment Text with Translation + WordNet
11*	Mohammad Soltani-2	0.8273	I+T	CLIP32	AdaBoost for Abortion + Xgboost for Gun Control
12*	Pitt Pixel Persuaders-2	0.8168	T		Emsemble All The Model
13	Mohammad Soltani-1	0.8142	I+T	CLIP32	AdaBoost for Abortion and Gun Control
14	Mohammad Soltani-4	0.8093	I+T	CLIP32	Xgboost for Abortion and Gun Control
15*	GC-HUNTER-2	0.8049	T	XLMLRoberta	
16	Mohammad Soltani-3	0.8000	I+T	CLIP32	AdaBoost for Abortion + RUSBoost for Gun Control
17	Pitt Pixel Persuaders-1	0.7910	T	BLOOM-560m	
18	Mohammad Soltani-5	0.7782	I+T	CLIP32	SVM-Poly for Abortion and Gun Control
19	GC-HUNTER-1	0.7766	T	BERT	
20*	IUST-1	0.7754	T+E	BERTweet	Augment Text with ChatGPT paraphraser + OCR on image
21	IUST-2	0.7752	T+E	RoBERTa	Augment Text with ChatGPT paraphraser + OCR on image
22	Pitt Pixel Persuaders-4	0.7710	T	Bloom-1B	
23	Pitt Pixel Persuaders-5	0.7415	T	XLNet	
24*	KPAS-1	0.7097	I+T	CLIP	
25*	ACT-CS-4	0.6325	I+T+E+C	ViT+BERT	Cross-Attention
26	ACT-CS-3	0.6178	I+T+E	ViT+BERT	Cross-Attention
27	ACT-CS-2	0.6116	I+T	ViT+BERT	Cross-Attention
28	ACT-CS-1	0.5863	I+T	ViT+BERT	Simple Concatenation of features
29	IUST-3	0.5680	I+T+E	CLIP+BERT	Augment Text with ChatGPT paraphraser + OCR on image
30	Pitt Pixel Persuaders-3	0.5285	I+T	ViLT	
31*	feeds-1**	0.4418	T	BERT	

Table 3: The Subtask-A submission results. The System column refers to the Team name and submission attempt number connected by "-". Each Team has at most five submissions. The scores are positive F1 scores. The T, I, E, and C represent text, image, extracted text from image, and image caption modality, respectively. Rows with **bold** ID and marked with * refer to the best system for each participating team. ** Team feeds submitted results for one topic by the submission deadline, so only partial results are evaluated.

4 Submission Results

We provide summaries about Subtask-A (Sec. 4.1) and Subtask-B (Sec. 4.2) submissions for all the teams. In cases where a team did not submit a description paper, we include their results and provide a brief description based on the survey completed by the team at the time of submission.

4.1 Subtask-A: AS Classification

Initially, we observed that models utilizing multimodal features (I+T or T+E) displayed higher performances, where I denotes tweet images, T denotes tweet text, and E denotes the text extracted from images. Table 3 illustrates that the top-performing submissions (top 10) employed two primary strategies: they either fused features extracted from both image and text encoders separately, or used pretrained language models finetuned on text extracted from images and tweets, which gave an additional textual context to the original tweet. This innovative method improved model performance compared to the ones that only used tweet text data in general⁶. Also, the last column shows that data augmentation exhibited promise, given the limited annotated data in this shared task.

4.1.1 System Descriptions

We describe representative methods from leading teams while summarizing the approaches from the remaining teams as follows:

Team KnowComp introduced a unified Framework for Text, Image, and Layout Fusion in Argument Mining, TILFA (Zong et al., 2023). They highlighted the need for better image encoding with textual information. To tackle the problem of unbalanced data, they augmented the tweet texts with backtranslation and synonym replacements.

Team Semantists (Rajaraman et al., 2023) submitted five system runs for task A, focusing mainly on the text-based approaches. To harness the information from the images, they extract text from the tweet image through an OCR system and concatenate it with the tweet texts. Pretrained language models such as T5 NLI (Raffel et al., 2020) and BERTTweet are applied for label predictions. The team also adopts a Multi-task Contrastive Learning Framework similar to Chen et al. (2022) with the label aware augmentation for contrastive learning.

⁶Results may vary depending on the model training details and experimental setups across participating teams

Team Mohammad Soltani (Soltani and Romberg, 2023) experimented with CLIP (Radford et al., 2021) to extract the textual and visual modality features. They then combined features from both modalities by concatenating them along the last dimension according to an early fusion strategy, followed by traditional machine learning classifiers such as AdaBoostClassifier and SVM-Poly.

Team Pitt Pixels Persuaders (Sharma et al., 2023) fine-tuned multiple text-based pre-trained models such as XLNet (Yang et al., 2019) and BLOOM (Scao et al., 2022) on the corpus. **Team IUUST** (Nobakhtian et al., 2023) did data augmentation using GPT to paraphrase tweet text and extracted text from images and finetuned text-based models. **Team feeds** (Torky et al., 2023) and **Team GC-Hunter** (Shokri and Levitan, 2023) only finetuned pre-trained language models on the tweet text. Both **Team ACT-CS** (Zhang et al., 2023) and **Team KPAS** studied multimodal feature fusions.

4.1.2 Method Discussions

Table 3 reveals that the most successful submissions utilized pretrained language models such as DeBERTa, BERT, and BERTTweet (Nguyen et al., 2020). Furthermore, the integration of data augmentation techniques, such as backtranslation and word substitution using WordNet, was observed to enhance performance, as depicted in Figure 2. This boost in performance can be attributed to the inherent reliance on textual information in the stance detection task. Augmenting the relatively limited annotated corpus with these techniques appears to be advantageous. Additionally, leveraging features from the visual modality, whether through image representations or image-text representations, further improved performance, ultimately leading to the highest overall scores, as demonstrated in Table 3 (rows 1 to 10).

On the other hand, the methods that utilized multimodal techniques like CLIP performed relatively lower than those that employed separate encoders for text and visual modalities. This is evident when referencing Table 3, where the system achieving the highest performance using CLIP as the joint encoder, namely the submission by Mohammad Soltani-2, is ranked 11th on the leaderboard. Additionally, it's noteworthy that only a limited number of teams explored the use of Large Language Models (LLMs). This might be attributed to our

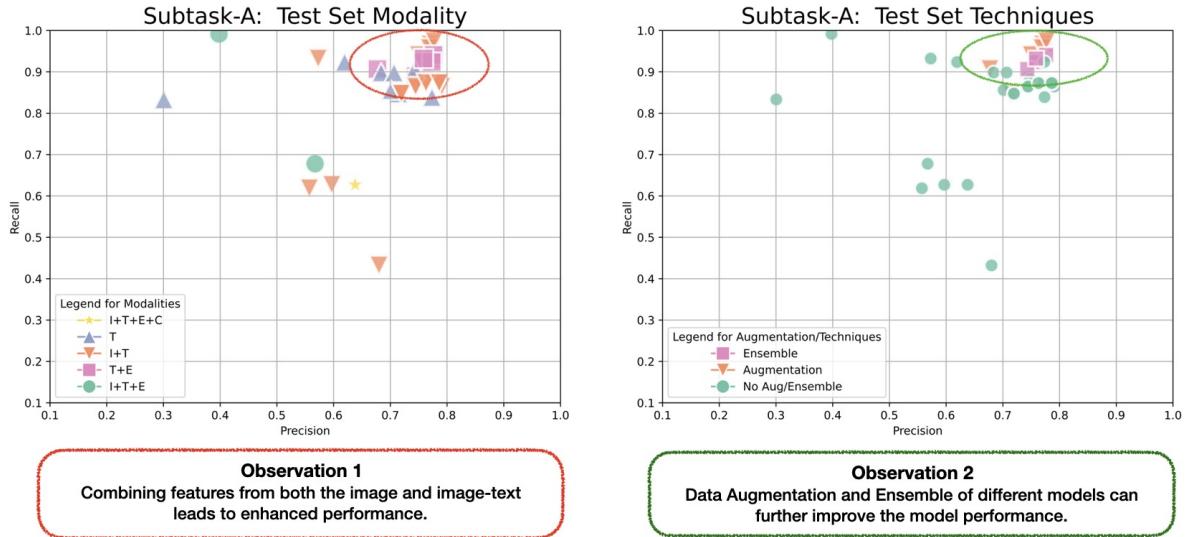


Figure 2: Subtask-A: system performance in relation to the computation approaches (left: modalities, right: techniques). We grouped systems based on the modalities used by the model (left) and computational techniques (right). The T, I, E, and C represent text, image, extracted text from image, and image caption modality, respectively.

initial guidelines⁷, which indicated that the utilization of commercial APIs like chatGPT⁸ would not contribute to the final ranking. Nevertheless, submissions that leveraged open-source LLMs, such as BLOOM-1B (row 22), exhibited lower performance compared to other submissions using pre-trained language models. This opens up opportunities for further research into exploring the capabilities of LLMs in understanding argumentation, especially in multimodal contexts.

4.1.3 Error Analysis

Figure 3 categorized the systems based on the modalities they incorporate and evaluated their respective success rates. Our analysis focused on system' ability to make accurate predictions, quantified by the number of successful systems out of 31 systems. We found that systems that incorporated both image and text modalities (I+T) generally yielded reasonable predictions, with at least one system in this category correctly identifying the label. Additionally, models that combined text and extracted text from images (T+E) displayed particularly strong performance, especially for data of intermediate difficulty. In these cases, the success rate for these systems exceeded 60%, with at least 19 out of the 31 systems making correct predictions.

In a qualitative analysis of the 299 valid tweets

⁷<https://imagearg.github.io/>

⁸<https://platform.openai.com/docs/guides/gpt/chat-completions-api>

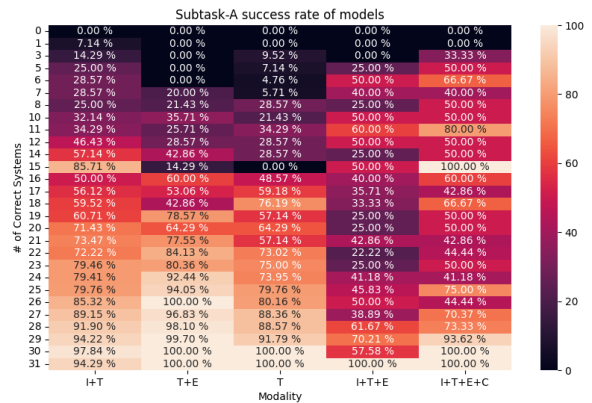


Figure 3: Average rate of correct predictions for Subtask-A systems (grouped by modalities) across tweet difficulties: the y-axis represents the number of systems making correct predictions out of 31 systems.

in the test set, we found that 160 tweets (53%) were accurately predicted by a majority of systems (≥ 26 out of 31 systems). Among the subset of tweets (86) exhibiting intermediate difficulty (where 6-20 teams failed to predict the correct labels), we manually sampled ten tweets for label analysis and provided potentially correct labels. Our findings indicate that these tweets often encompass cynicism or sarcasm regarding a specific topic (3 cases), are heavily reliant on the image contents/charts (3 cases), or can be traced back to annotation noise or contents unrelated to the provided topic. Detailed insights are shown in Table 5 in Appendix A. For instance, the first example associates "pro-life" with "Abortion Law", suggesting the tweets favor

abortion. In the second example, a deep understanding of the text embedded within images is crucial for providing accurate labels. These observations underscore the complexities in multimodal argument mining tasks and highlight the critical role of cross-modal information fusion.

4.2 Subtask-B: IP Classification

In contrast to Subtask-A, participating teams made fewer submission attempts for Subtask-B (a total of 21 compared to 31 for Subtask-A). Notably, all submissions in Subtask-B employed approaches that incorporated multiple modalities, as this task inherently requires an integration of visual and textual information to assess image persuasiveness.

As shown in Table 4, utilizing CLIP (Radford et al., 2021) model is evident to be the most effective technique in extracting multimodal features, which yields the best results (top-4 systems leveraged CLIP). This indicates that a unified encoder can better model the cross-modal information fusion, compared to employing individual models (i.e., ViT (Dosovitskiy et al., 2020) for image and BERT (Devlin et al., 2019) for text) for feature extractions. Moreover, three teams utilized off-the-shelf Optical Character Recognition (OCR) tools to extract image text content. This extracted text was then combined with the original tweet texts to fine-tune pre-trained language models, which suggests that users could include arguments through texts embedded in the images.

4.2.1 System Descriptions

We describe systems from the top-performing teams and briefly summarize the remaining teams:⁹

Team feeds (Torky et al., 2023) made 2 submissions (Table 4 rows 1 and 3). The team utilized the CLIP model to encode the image and text and use a simple concatenation to fuse the two modalities, then trained a neural network on the concatenated features. They carefully cleaned tweet texts by recovering common abbreviations with their full forms (such as "I'm to I am") and also removed content such as URLs, emails, and phone numbers.

Team KPAS did not submit a system demonstration paper. However, their submission notes showed that they also employed the CLIP model to extract multimodal features.

Team Mohammad Soltani (Soltani and Romberg, 2023) made a total of 5 submissions

⁹While Team KPAS was among the top-performing teams, they did not submit a system description paper.

(Table 4 rows 4, 7, 8, 9, and 12). Notably, they adopted a topic-specific approach, tailoring their strategies to each topic separately. For the "Abortion" topic, they integrated visual features extracted from the CLIP model and utilized them as inputs for a classifier. Conversely, when tackling the "gun control" topic, their most successful model was crafted by combining features from Reformer (Kitaev et al., 2019), ELECTRA (Clark et al., 2019), and LayoutLM (Xu et al., 2020).

Similar to the systems in Subtask-A, **Team Semantists** (Rajaraman et al., 2023) extracted texts from images and fine-tuned pretrained Language models such as T5 NLI and StancyBERT (Popat et al., 2019) on the corpus. **Team ACT-CS** (Zhang et al., 2023) and **Team KnowComp** (Zong et al., 2023) used separate models to encode the visual and textual information individually, then fine-tuned classifiers based on the fused features. **Team IUST** (Nobakhtian et al., 2023) (Table 4 row 11) leveraged the MultiModal Bit Transformer to extract features from both image and text sources concurrently. **Team GC-Hunter** (Shokri and Levitan, 2023) chose to concatenate text content from both tweets and OCR outputs to fully leverage textual information, complemented by image features extracted from a separately trained ViLT model. Finally, **Team Pitt Pixel Persuaders** (Sharma et al., 2023) (Table 4, row 21) did not include the details of their Subtask B submission in their system description paper. However, their submission notes reveal that they also relied on CLIP, which proved to be less successful in their case.

4.2.2 Method Discussion

Figure 4 illustrates that, unlike Subtask A, the application of data augmentation techniques which primarily concentrated on augmenting the text modality exclusively obtained only modest improvements in classification performance. Notably, none of the participating teams explored augmentation for the visual modalities, which presents an opportunity for further research into the impact of image augmentation on enhancing persuasiveness detection.

Additionally, Table 4 indicates that none of the submissions integrated LLMs into their systems. This observation can also be attributed to the task's primary emphasis on both visual and textual modalities and the guidelines we enforced, which limited the use of LLMs to open-source models. These open-source models have received less attention within the context of multimodal tasks, providing

ID	System	Scores	Modality	Model	Notes
1*	feeds-1	0.5561	I+T	CLIP	Cleaned Text
2*	KPAS-2	0.5417	I+T	CLIP	
3	feeds-2	0.5392	I+T	CLIP	Uncleaned Text
4*	Mohammad Soltani-5	0.5281	I+T	CLIP32+REL+Convnext	
5*	Semantists-1	0.5045	T+E	T5	OCR on Image
6*	ACT-CS-1	0.5000	I+T	Vit+BERT	
7	Mohammad Soltani-1	0.4875	I+T	CLIP32	SVM-Poly for Abortion LogisticReg for Gun Control
8	Mohammad Soltani-4	0.4778	I+T	CLIP32+REL+Convnext	SGD for Abortion LogisticReg for Gun Control
9	Mohammad Soltani-3	0.4762	I+T	CLIP_L_14	SVM-Poly for Abortion and Gun Control
10	Semantists-5	0.4659	T+E		Emsemble with majority vote
11*	IUST-1	0.4609	I+T	CLIP+BERT	Augment Text with ChatGPT paraphraser + OCR on image
12	Mohammad Soltani-2	0.4545	I+T	CLIP32	SGD for Abortion and Gun Control
13	ACT-CS-4	0.4432	I+T+E+C	Vit+BERT	Cross Attention
14	ACT-CS-3	0.4348	I+T+E	Vit+BERT	Cross Attention
15	Semantists-4	0.4222	T+E		Emsemble with consistency loss
16	Semantists-2	0.4141	T+E	Stancy BERT	
17*	KnowComp-1	0.3922	I+T	LayoutLMv3+DeBERTa	Augment Text with Translation + WordNet
18*	GC-HUNTER-1	0.3832	I+T+E	ViLT	OCR on Image
19	ACT-CS-2	0.3125	I+T	Vit+BERT	Cross Attention
20	Semantists-3	0.2838	I+T+E	ALBEF	
21*	Pitt Pixel Persuaders-1	0.1217	I+T	CLIP	

Table 4: The Subtask-B submission results. Each Team is allowed at most 5 submissions. The scores are positive label F1. The T, I, E, and C, represent text, image, extracted text from image, and image caption modality, respectively. Rows with **bold** ID and marked with * refer to the best system for each participating team.

an explanation for their absence in the submissions.

4.2.3 Error Analysis

Figure 5 categorizes the systems based on the modalities they incorporate and their respective success rates. Our analysis focused on the models’ ability to make accurate predictions, quantified by the number of successful systems out of the 21 total systems. We found that systems incorporating both image and text modalities (I+T) consistently produced accurate predictions across data points with varying levels of difficulty. Interestingly, systems that combined text, text on images, images, and captions (I+T+E+C) demonstrated strong performance, particularly for data with high difficulty levels (as indicated by rows where only 4/5 systems made correct predictions). As reported by [Soltani and Romberg \(2023\)](#), these systems tended to classify

images showing only text as persuasive. Further analysis on the data illustrated different argumentation techniques, such as cases, consequences, or outcomes related to the textual argument, further highlighting the complexity and diversity of approaches employed in this shared task.

5 Conclusion

In this paper, we introduced the *ImageArg* shared task, marking a significant milestone as the inaugural shared task in multimodal argument mining, co-located with the 10th Argument Mining Workshop at EMNLP 2023. A total of 9 teams from 6 different countries enthusiastically participated in this task, collectively submitting 31 systems for Subtask-A Argument Stance (AS) classification and 21 systems for Subtask-B Image Persuasiveness (IP) classification. The results reveal that

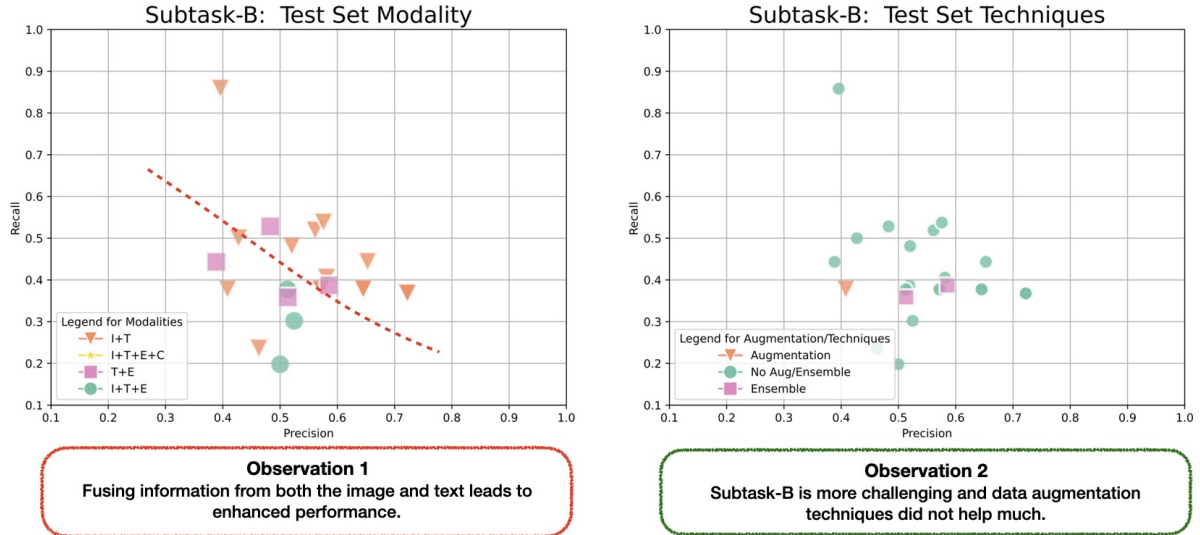


Figure 4: Subtask-B: system performance in relation to the computation approaches (left: modalities, right: techniques). We grouped systems based on the modalities used by the model (left) and computational techniques (right). The T, I, E, and C represent text, image, extracted text from image, and image caption modality, respectively.

Subtask-A is comparatively more predictable than Subtask-B. Models that utilized both textual information and the text embedded within images demonstrated considerable performance in Subtask-A. Furthermore, the strategic use of data augmentation and ensemble methods further enhanced the models’ effectiveness. In contrast, Subtask-B witnessed the predominant adoption of CLIP for feature extraction from both images and texts, a technique that exhibited significant promise. The two subtasks offered valuable opportunities for participants to actively engage and foster fruitful exchanges in multimodal argument mining research.

6 Limitations

In this section, we discuss the limitations of our work from multiple perspectives. First, the datasets utilized in this task may not sufficiently cover a broad range of multimodal data, possibly leaning toward social media content related to two specific topics: gun control and abortion. The language of data included in the paper is English, which is limited and should be extended to other languages for argument mining. Meanwhile, as demonstrated in Section 4.1.3, the label annotations may exhibit inconsistencies or inaccuracies, given the inherent complexity of the task. Also, the use of rhetorical devices, especially in addressing challenges like sarcasm detection, remains an underexplored area. The evaluation metrics employed may not fully encompass the nuanced performance aspects crucial

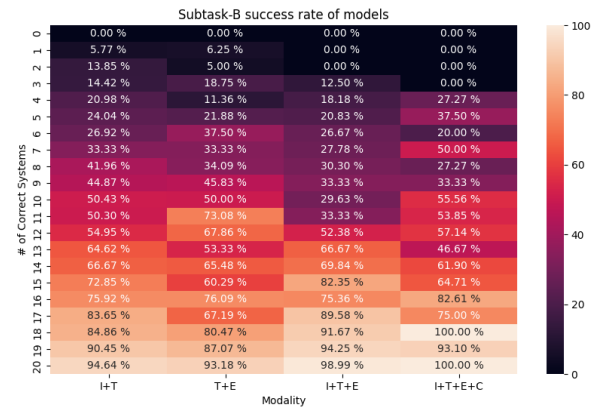


Figure 5: Average rate of correct predictions for Subtask-B systems (grouped by modalities) across tweet difficulties: the y-axis represents the number of systems making correct predictions out of 21 systems.

for multimodal argument mining. Lastly, it’s important to acknowledge that participating systems may encounter challenges when attempting to generalize their approaches across diverse data types, domains, or modalities.

Regarding the analysis of the results, it’s important to acknowledge that since we mainly collected final predictions for both subtasks, the interpretability of the systems might remain unclear, presenting challenges in gaining insights into their decision-making processes. The intricate nature of multimodal argument mining can lead to multiple valid interpretations, potentially affecting the clarity of the ground truth.

7 Ethics

We acknowledge that there are privacy and ethical considerations in the collection and utilization of social media data. It's possible that biases within the dataset or system outputs may not have been fully mitigated. Given that our data originates from Twitter and the annotators predominantly come from English-speaking countries, it's inevitable that cultural biases are inherent in the data. However, we have implemented several measures to mitigate potential risks. To address privacy concerns, we have chosen to publicly share only the tweet IDs with the research community, which aligns with Twitter Developer Policy¹⁰.

References

- Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. **Findings of the third shared task on multimodal machine translation**. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.
- Winston Carlile, Nishant Gurrupadi, 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.
- Miriam Louise Carnot, Lorenz Heinemann, Jan Braker, Tobias Schreieder, Johannes Kiesel, Maik Fröbe, Martin Potthast, and Benno Stein. 2023. On stance detection in image retrieval for argumentation.
- Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen 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.
- 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.
- 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*.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.
- 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. In *International Conference on Learning Representations*.
- Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument: Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*, pages 299–310. IOS Press.
- 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.
- Meiqi Guo, Rebecca Hwa, and Adriana Kovashka. 2021. Detecting persuasive atypicality by modeling contextual compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 972–982.
- Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.
- Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting forum*, volume 36, pages 194–208. Elsevier.
- Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. 2019. Dense multimodal fusion for hierarchically joint representation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3941–3945. IEEE.
- Xinyue Huang and Adriana Kovashka. 2016. Inferring visual persuasion via body language, setting, and deep features. *2016 IEEE Conference on Computer*

¹⁰<https://developer.twitter.com/en/developer-terms/policy>

- Vision and Pattern Recognition Workshops (CVPRW)*, pages 778–784.
- 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.
- Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.
- 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.
- John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- 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.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.
- 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.
- 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.
- Melika Nobakhtian, Ghazal Zamaninejad, Erfan Moosavi Monazzah, and Sauleh Eetemadi. 2023. Just at imagearg: 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.
- Daniel J O’Keefe. 2015. *Persuasion: Theory and research*. Sage Publications.
- Sunghyun Park, Han Suk Shim, Moitreyia 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*, pages 50–57.
- Isaac Persing and Vincent Ng. 2017. Why can’t you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4082–4088.
- Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. Stancy: Stance classification based on consistency cues. 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 6413–6418.
- 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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- 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. *Journal of Machine Learning Research*, 21(140):1–67.
- Kanagasabai Rajaraman, Hariram Veeramani, Saravanan Rajamanickam, Adam Maciej Westerski, and Jung-Jae Kim. 2023. Semantists at imagearg-2023: Exploring cross-modal contrastive and ensemble models for multimodal stance and persuasiveness classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Matilda White Riley. 1954. Communication and persuasion: psychological studies of opinion change.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon,

- Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Arushi Sharma, Abhibha Gupta, and Maneesh Bilalpur. 2023. Argumentative stance prediction: An exploratory study on multimodality and few-shot learning. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Mohammad Shokri and Sarah Ita Levitan. 2023. Gc-hunter at imagearg shared task: Multi-modal stance and persuasiveness learning. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Mohammad Soltani and Julia Romberg. 2023. A general framework for multimodal argument persuasiveness classification of tweets. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. [A shared task on multimodal machine translation and crosslingual image description](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.
- Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.
- Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *ArgNLP*, pages 21–25.
- Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. 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 5103–5114.
- Islam Torky, Simon Ruth, Shashi Sharma, Mohamed Salama, Krishna Chaitanya, Tim Gollub, Johannes Kiesel, and Benno Stein. 2023. Team feeds @ imagearg 2023: Embedding-based stance and persuasiveness classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. In *International Conference on Learning Representations*.
- 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.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. [Layoutlm: Pre-training of text and layout for document image understanding](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 1192–1200, New York, NY, USA. Association for Computing Machinery.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.
- Jing Zhang, Shaojun Yu, Xuan Li, Jia Geng, Zhiyuan Zheng, and Joyce Ho. 2023. Split: Stance and persuasion prediction with multi-modal on image and textual information. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Hanzhong Zheng, Zhexiong Liu, and DeJia Shi. 2021. [Image-text discourse coherence relation discoveries on multi-image and multi-text documents](#). *Journal of Physics: Conference Series*, 1948(1):012013.
- Qing Zong, Zhaowei Wang, Baixuan Xu, Tianshi Zheng, Haochen Shi, Weiqi Wang, Yangqiu Song, Ginny Wong, and Simon See. 2023. Tilfa: A unified framework for text, image, and layout fusion in argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

A Appendix



Image	Text	Annotations
	<p>'Abortion law is pro-life. It saves 'mother over 'growing fetus in unwanted pregnancy due to rape, psychological trauma, social stigma, etc. It stops back-alley abortions that kill. Counseling & transition homes can lessen 'need for abortion.</p>	<p>Topic: Abortion Annotated Label: Oppose System Predictions: {'Oppose': 19, 'Support':12} Potentially Correct Label: Support Rationale: The human annotation is inaccurate, super interesting on the usage of 'pro-life', to advocate for abortion.</p>
	<p>How Pro-Life is the Republican party and Justices? Facts matter here the answer, they're not. Thanks to their rulings, women have been able to safely have abortions. #RoeVWade #Republicans #SCOTUShearings #Constitution #prochoice #ProLife #Facts</p>	<p>Topic: Abortion Annotated Label: Support System Predictions: {'Oppose': 20, 'Support':11} Potentially Correct Label: Support Rationale: This tweet uses sarcasm, and is hard to annotate (republicans are in general not supporting legal abortion). Here the contents are image-dependent.</p>

Table 5: Manually checked data with controversial scenarios for Subtask-A, where nearly half of the systems failed to predict the correct label. We sampled a few tweets and provided a potential correct label based on our manual inspections. The first example redefines a widely used anti-abortion term, pro-life, and advocates for abortion instead. The second is a complicated one that requires the comprehension of texts embedded in the image.