SPLIT: Stance and Persuasion Prediction with Multi-modal on Image and Textual Information

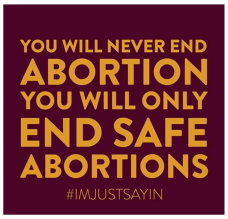
Jing Zhang[1], **Shaojun Yu** [1], **Xuan Li** [2], **Jia Geng**[3], **Zhiyuan Zheng**[4], **Joyce C Ho**[1]
[1] Emory University [2] Carnegie Mellon University
[3] University of Miami [4] American Cancer Society
{jing.zhang2, shaojun.yu, joyce.c.ho}@emory.edu
xuanli1@andrew.cmu.edu jxg570@miami.edu jason.zheng@cancer.org

Abstract

Persuasiveness is a prominent personality trait that measures the extent to which a speaker can impact the beliefs, attitudes, intentions, motivations, and actions of their audience. The ImageArg task is a featured challenge at the 10th ArgMining Workshop during EMNLP 2023, focusing on harnessing the potential of the ImageArg dataset to advance techniques in multimodal persuasion. In this study, we investigate the utilization of dual-modality datasets and evaluate three distinct multi-modality models. By enhancing multi-modality datasets, we demonstrate both the advantages and constraints of cutting-edge models.

1 Introduction

Persuasion encompasses the art of one party endeavoring to influence another's thoughts, beliefs, or actions, and it stands as a fundamental and versatile human capability. Its significance goes far beyond the realms of business and politics, permeating numerous facets of our everyday existence. In the fast-changing realm of natural language processing (NLP) and artificial intelligence (AI), there has been a notable increase in enthusiasm for creating techniques and datasets to enhance and assess persuasiveness in natural language applications (Hunter et al., 2019; Chatterjee and Agrawal, 2006; Liu et al., 2022). The capacity to convince, sway, and captivate using language has long been a fundamental element of human communication, and with the emergence of advanced language technologies, the pursuit of leveraging persuasive capabilities in digital interactions has gained remarkable momentum. In today's digital age, the proliferation of social media platforms has ushered in a new frontier for the practice of persuasion. These platforms serve as fertile ground, affording both organizations and individuals the opportunity to engage in activities that extend beyond mere persuasion and can include disinformation campaigns.



Which will only cause more harm to both the woman and the precious fetus you want to save but won't take care of. thank you for proving once again u r pro/forced birthers; hate women. #mybodymychoice #abortion #prochoice #prolife #hypocrites #childfree #abortionlaw #AbortionBan

Stance: Support
Pervasiveness: Yes

Figure 1: The abortion tweet picture (left) and its tweets (right) from Liu et al. (2023).

The pervasive reach and influence of social media amplify the potential impact of persuasive efforts, making it imperative for individuals and society as a whole to exercise discernment and critical thinking in navigating this dynamic landscape.

Most of current works in argumentation mining solely focus on textual format, such as the argumentation dialogues (Hunter et al., 2019), contextual advertising (Wen et al., 2022), and other works (Lukin et al., 2017; Persing and Ng, 2017). In their work, Nojavanasghari et al. (2016) introduced a comprehensive deep multimodal fusion approach to predict persuasiveness, incorporating three modalities: Visual, Acoustic, and Text. Nevertheless, in light of the current trend observed on Twitter, as depicted in Figure 1, it becomes evident that numerous images accompanied by text are surfacing. Mere application of computer vision (CV) techniques for object recognition proves inadequate for addressing this challenge. Liu et al. (2022) designed two tasks based on the tweets, Stance detection and Persuasion prediction. Stance detection (SD) involves the automated task of ascertaining, based on textual content, whether the author expresses a supportive, opposing, or neutral position regarding a particular proposition or subject. This subject can encompass individuals, organizations, government policies, movements, products, and more. As an illustration, considering the tweet and accompanying image in Figure 1, it is evident that the stance conveyed is one of support. Persuasion prediction (PP) determines the degree of

Proceedings of the 10th Workshop on Argument Mining, pages 175–180

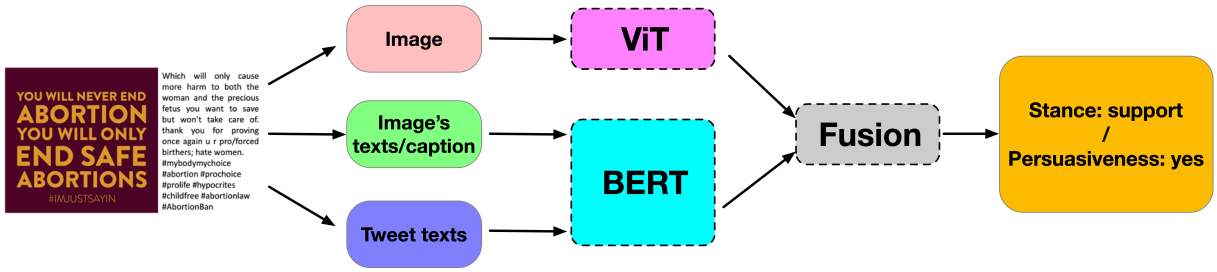


Figure 2: The overview of SPLIT framework.

persuasiveness or the potential impact that a given tweet may have on its readers or the broader audience. Given the unique characteristics of the existing Twitter data, this paper will design additional feature extraction methods, such as using Optical Character Recognition (OCR) to extract text from images, in order to enrich the feature space. This will enable a more comprehensive analysis of the stance and persuasion in the current tweets. Our code is publicly available in GitHub (https://github.com/JZCS2018/ACT-CS). In summary, our contributions are as follows:

- We combine current state-of-the-art (SOTA) CV and NLP models as SPLIT, to utilize the image and textual information for the SD and PP tasks.
- We align the individual tweet's text, image, and its textual information (texts in image and generated image caption), and utilize different fusion methods to show the detailed analysis.

2 Related Works

Persuasiveness Prediction Persuasiveness prediction is an under-explored topic but has attracted growing interests(Chatterjee et al., 2014; Park et al., 2016; Lukin et al., 2017; Carlile et al., 2018; Chakrabarty et al., 2020). As the majority of works (Higgins and Walker, 2012; Lukin et al., 2017; Persing and Ng, 2017; Carlile et al., 2018) utilized textual inputs - such as audience variable, report, and student essays - to analyze persuasion strategies, (Joo et al., 2014; Huang and Kovashka, 2016) pioneered the study of in persuasion in social media with visual information, including facial expression, body gesture and human portrait. Finally, Hussain et al. (2017); Guo et al. (2021) investigated sentiment, intent reasoning and persuasive strategies in advertisement context in multi-modal learning. However, a persuasive-targeted and multi-modal framework is still missing in the current NLP literatures.

Multi-modal learning Thanks to the progress in language models and alignment techniques, multi-modal learning with text and image have recently received significant attention in the CV and NLP communities. As the majority of SOTA works are built upon transformersand and its variants, different alignment strategies have been proposed and applied to fuse representations from each modality. On the one hand, many works (Radford et al., 2021; Neelakantan et al., 2022) employ modality-specific encoders and apply contrastive loss to align representations. The encoders (Dosovitskiy et al., 2020; Devlin et al., 2018) are usually pretrained to learn visual and textual representations independently and kept frozen during alignment. On the other hand, many recent works (Bao et al., 2022; Li et al., 2023a; Zhang et al., 2023; Sun et al., 2023; Koh et al., 2023) have tokenized visual representations and grounded them to unifed language model for multimodal tasks. Specifically, the visual and textual tokens are concatenated as input to the pre-trained language model, and then are aligned through various tasks such as next token prediction. However, multi-modal learning on stance and persuassive prediction are under-explored, partially due to a lack of multi-modal corpora and persuasive-specific modeling framework.

3 Approach

Let D be a tweet dataset, where each tweet d_i is represented as a tuple (I_i, T_i). I_i represents the image associated with the tweet and T_i represents the textual content of the tweet. A model f that maps the input tuples (I_i, T_i) to a predicted stance score or persuasiveness score $\hat{y}_i$, where $\hat{y}_i \in [0, 1]$. In addition, we will also extract the text in the image It_i, and generate its caption C_i as an optional feature. Then the representation of the tweet can be shown

as the new tuple $(I_i, T_i, It_i/C_i)$. The framework is illustrated in Figure 2.

3.1 OCR & image captioning

Due to the limited amount of training data available, we believe that incorporating other pretrained models will significantly enhance the performance of our model. Therefore, we have incorporated two types of pretrained models in our approach: BLIP-large (Li et al., 2022), an image captioning model for generating textual descriptions, and Microsoft's TrOCR(Li et al., 2023b), an optical character recognition (OCR) model for extracting text from images. BLIP-large has been pre-trained on a vast dataset and is capable of generating textual descriptions for images. By utilizing this model, we aim to improve the understanding and contextual description of the images in our dataset. Additionally, some images contain text that is crucial to comprehend the image but the text cannot be effectively represented solely through captions especially for longer texts. To address this, we employ the OCR model to extract text from these images.

For each image I_i, we use BLIP-large model to generate the caption C_i and use TrOCR to extract the text It_i. These two features are then directly fed into our backbone model.

3.2 Backbone Models

As the fields of CV and natural NLP continue to advance, we aim to integrate SOTA models from both domains for our tasks.

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) is designed for CV tasks, and it offers several compelling benefits for the image processing. Its ability to extract intricate visual patterns and characteristics from images has demonstrated remarkable effectiveness. It can seamlessly integrate with other models, especially the BERT (Devlin et al., 2018) for text, enabling the creation of powerful multi-modal models.

BERT is pre-trained on vast amounts of text data and has a deep understanding of contextual language usage. This makes it highly effective in capturing nuanced language patterns and context within tweets, which is crucial for analyzing persuasiveness.

The self-attention mechanisms in ViT and BERT models could provide insights into which parts of the image/text the model focuses on when making predictions. This interpretability can be valuable for understanding how the model assesses the stance and persuasiveness.

3.3 Fusion Methods

Multimodal fusion methods are techniques used to combine and integrate information from multiple modalities (e.g., text, images, audio) into a unified representation for analysis or decision-making (Gao et al., 2020). It can be categorised into early fusion, late fusion and intermediate fusion. Early fusion, also known as feature-level fusion, involves combining features from different modalities at the input level. For example, in text-image fusion, the features extracted from text and images are concatenated or merged before being fed into a model. This approach creates a single feature vector that represents both modalities. Late fusion, involves processing each modality separately and then combining their results at a later stage. Cross-attention was introduced in Transformers model (Vaswani et al., 2017). It often employs attention mechanisms to enable a model to selectively attend to relevant parts of one modality based on the information from another modality. This paper will apply the three methods to our experiments.

4 Experiments

We designed the experiments to answer two key questions: (1) How *accurate* is `SPLIT` in automating the entity matching? (2) How *important* are the different components of `SPLIT`?

4.1 Datasets

The benchmark dataset used in this study is sourced from the ImageArg-Shared-Task-2023, as described in Liu et al. (2023). This dataset encompasses two specific topics: abortion and gun control. In the abortion dataset, there are 891 training samples, 100 validation samples, and 150 test samples. Similarly, the gun control dataset comprises 923 training samples, 100 validation samples, and 150 test samples. For each topic, we will experiment on stance and persuasiveness prediction tasks.

4.2 Baseline models

We utilize the pretained ViT and BERT-based-uncased models for our experiments. To ensure a fair comparison, we standardize the dimensionality of both image and text embeddings to 1024 before inputting them into the classification layers. We evaluate task performance across three modalities: Image Modality (I-ViT), Text Modality (T-BERT),

Datasets	Tasks	I-ViT	T-BERT	SPLIT-IT-E	SPLIT-IT-L	SPLIT-IT	SPLIT-IET	SPLIT-IECT
Total	Stance	0.4279	0.4738	0.5863	0.6098	0.6116	0.6178	**0.6325**
	Persuasiveness	0.3968	0.3906	**0.5000**	0.4076	0.3125	0.4348	0.4432
Abortion	Stance	0.3609	0.3975	0.4337	0.4429	0.4595	0.4494	**0.4638**
	Persuasiveness	0.5438	0.4751	**0.605**	0.5982	0.3333	0.4950	0.4510
Gun control	Stance	0.4782	0.5315	0.6627	0.6689	0.6786	0.7059	**0.7030**
	Persuasiveness	0.2192	0.2908	0.3529	0.3017	0.2895	0.3614	**0.4337**

Table 1: Comparison of F1 performance for different models. The best performance is bolded.

and Multi-modality combining both text and image information. For last part, we try different configurations, such as Image + Text + Early fusion (SPLIT-IT-E), Image + Text + Late fusion (SPLIT-IT-L), Image + Text + Cross-attention (SPLIT-IT) Image + Text-extraction + Text + Cross-attention (SPLIT-IET), and Image + Text-extraction + Image-caption + Text + Cross-attention (SPLIT-IECT).

We train all models on a single NVIDIA Tesla V100 GPU with 16GB VRAM. We fix the batch size at 32 and use the Adam optimizer to train the models for 20 epochs using a linearly decaying learning rate with one epoch warmup. A learning rate sweep is done over the range [1e-5, 3e-5, 5e-5, 8e-5. 1e-4]. We also apply the early stopping strategy for the efficiency.

5 Results

5.1 Predictive Performance on Different Tasks

The Table 1 shows the results from different models on different datasets and tasks. The total datasets means we only consider the tasks instead of topics for the evaluation. For the "Stance" task in the "Total" dataset, "SPLIT-IECT" achieves the highest F1 score of 0.6325, making it the best-performing model. Among single-modality models, T-BERT outperforms I-ViT, indicating that text holds a more significant role in this Stance task. When considering the outcomes of multi-modal models, it becomes evident that incorporating text information extracted from images has a positive impact on model performance. In the context of the "Persuasiveness" task, "SPLIT-IT-E" emerges as the top-performing model, achieving an F1 score of 0.5000. Despite observing improved performance with the incorporation of additional features, it appears that the inclusion of textual information does not significantly contribute to enhancing the decision-making process. This also can be observed in the comparison between I-ViT and T-BERT.

5.2 Predictive Performance on Different Topics

In the "Abortion" topic, "SPLIT-IECT" again performs the best for the "Stance" task with an F1 score of 0.4638. However, for the "Persuasiveness" task, "SPLIT-IT-E" has the highest F1 score of 0.605. The textual content within the images is evidently more pivotal in aiding the decision-making process. Furthermore, the outcomes in the Persuasiveness task align consistently with those observed in the overall dataset for the same task.

In the context of the "Gun control" topic, "SPLIT-IECT" takes the lead in the "Stance" task, achieving an F1 score of 0.7030. Similarly, in the "Persuasiveness" task within the same topic, "SPLIT-IECT" maintains its superior performance with an F1 score of 0.4337. Notably, the results in this particular topic differ from those observed in other topics. It appears that the images within the Gun control dataset contain more valuable textual information compared to those in the Abortion dataset."

Finally, when examining fusion techniques, it becomes evident that cross-attention mechanisms can offer more potent insights for predicting outcomes.

6 Conclusion

In light of the recent advancements in persuasiveness and stance prediction research, this study combines state-of-the-art computer vision (CV) and natural language processing (NLP) models under the name SPLIT, and explores various fusion approaches. The findings indicate that the cross-attention mechanism outperforms other methods. In the future, we will focus on how to visualize and interpret the predictions from the model, which could provide more comprehensive analysis to the researchers.

Acknowledgements. This work was supported by the National Science Foundation award IIS-2145411.

References

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

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

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

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.

Niladri Chatterjee and Saumya Agrawal. 2006. Word alignment in english-hindi parallel corpus using recency-vector approach: some studies. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 649–656.

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*.

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

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

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.

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.

Xinyue Huang and Adriana Kovashka. 2016. Inferring visual persuasion via body language, setting, and deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 73–79.

Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*, pages 18–33. Springer.

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*, pages 1705–1715.

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.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. *ICML*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023b. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

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.

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

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

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

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2016. Multimodal analysis and prediction of persuasiveness in online social multimedia. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(3):1–25.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

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.

Taylor Jing Wen, Ching-Hua Chuan, Jing Yang, and Wanhsiu Sunny Tsai. 2022. Predicting advertising persuasiveness: A decision tree method for understanding emotional (in) congruence of ad placement on youtube. *Journal of Current Issues & Research in Advertising*, 43(2):200–218.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*.