Webis @ ImageArg 2023: Embedding-based Stance and Persuasiveness Classification

Islam Torky[1] Simon Ruth[1] Shashi Sharma[1] Mohamed Salama[1]

Krishna Chaitanya[1] Tim Gollub Johannes Kiesel Benno Stein

Bauhaus-Universität Weimar, Germany

Abstract

This paper reports on the submissions of Webis to the two subtasks of ImageArg 2023. For the subtask of argumentative stance classification, we reached an F1 score of 0.84 using a BERT model for sequence classification. For the subtask of image persuasiveness classification, we reached an F1 score of 0.56 using CLIP embeddings and a neural network model, achieving the best performance for this subtask in the competition. Our analysis reveals that seemingly clear sentences (e.g., "I support gun control") are still problematic for our otherwise competitive stance classifier and that ignoring the tweet text for image persuasiveness prediction leads to a model that is similarly effective to our top-performing model.

1 Introduction

In recent years, the analysis of the argumentative stance of images and texts has gained significant attention. Several shared tasks have been conducted in this area, like the same-side stance classification (Körner et al., 2021) on texts, and the image retrieval for arguments (Bondarenko et al., 2022, 2023) on images. However, especially for images, the task of stance detection is far from being solved (Carnot et al., 2023). The ImageArg 2023 competition then provided a platform for researchers to explore this task further in the multi-modal context of tweets with images. Moreover, the competition featured a second task of predicting whether the image enhanced the persuasiveness of the text.

In this paper, we present the work conducted by our team, "feeds," for the ImageArg 2023 competition. Our efforts led to insightful findings and promising results in both tasks, shedding light on the complexities of combining visual and textual information for argumentative analysis.

For subtask A (argumentative stance classification), we employed a BERT model (Devlin et al., 2019) with stacked Transformer encoders. A separate model was trained for each of the two topics. Training encompassed tokenization, batch processing, optimizer, and learning rate optimization for F1 scores on the validation set. Our approach achieved an F1 score of 0.84 on the test set.

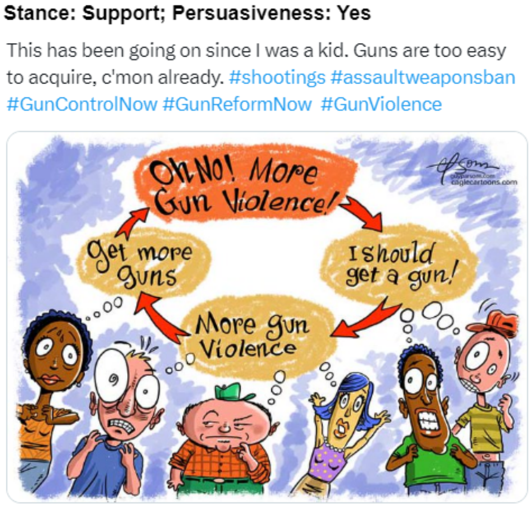
For subtask B (image persuasiveness classification), we employed the CLIP model (Radford et al., 2021) and a linear neural network. We integrated image and text embeddings to have multimodal features fed into the neural network. Tests with separate models and combined models for the two tasks were conducted. When removing the text features, we still get similar performance compared to using both features. Therefore image features seem more decisive for persuasiveness than the text and the multimodality of this task is hard to leverage. We achieved an F1 score of 0.56 on the test set, which is the highest among all submissions.

This paper is structured as follows: Section 2 provides a brief overview of related work. In Section 3, we detail our methodology and approaches for both subtasks. Section 4 presents our results and their implications, while Section 5 discusses the obtained results. Finally, Section 6 concludes the paper, summarizing our contributions and outlining potential directions for future research.

2 Related Work

Argumentative stance detection is still considered a major problem in NLP. Ajjour and Al-Khatib (2021) analyzed several stance classifiers for textual arguments, which achieved an accuracy between 0.50 to 0.77, and identified as challenges an inadequate topic knowledge of classifiers or when arguments only partial agree or disagree. Similarly, Carnot et al. (2023) identified several challenges for detecting the stance expressed in images when analyzing the submissions to the Touché 2022 shared task on image retrieval for argumentation (Bondarenko et al., 2022): bridging the seman-

[1] Authors contributed equally



Stance: Support; Persuasiveness: Yes

This has been going on since I was a kid. Guns are too easy to acquire, c'mon already. #shootings #assaultweaponsban #GunControlNow #GunReformNow #GunViolence

Figure 1: Example of tweet from the dataset, showing support for gun control and with the image increasing the persuasiveness of the text (class=yes).

tic gap for diagrams, ambiguity arising from diverse valuations leading to varied interpretations, the dependence of image understanding on background knowledge, regional relevance, the presence of both stances in one image, irony, and more. All of these also apply here, but maybe to a lesser degree as classifiers were trained for each topic. Liu et al. (2022) dealt with multi-modal analysis in persuasiveness classification. They identified an issue that the image encoder could not capture text like slogans in images. They suggested extracting and using textual features from images.

3 Task

We participated in both ImageArg subtasks:

Subtask A: Argumentative Stance Classification. Given a tweet with text and an image, predict if the tweet supports or opposes a topic.

Subtask B: Image Persuasiveness Classification. Given a tweet with text and an image, predict if the image makes the tweet text more persuasive.

For both subtasks, the organizers provide a human-annotated dataset of 2K tweets (Liu et al., 2022).[2] Submissions are evaluated using F1 score. For illustration, Figure 1 shows an example tweet for the gun control topic with associated classes: "support" for subtask A and "yes" for subtask B.

4 Our Approach

We employed neural models on text and image embeddings for tackling the tasks. For training, we either trained two *separate* models for the two topics of the dataset ("gun control" and "abortion") to capture topic-specific characteristics, or trained a *combined model* on both topics to capture topic-independent features. We then describe data preprocessing (Section 4.1), and the models used in subtask A (Section 4.2) and B (Section 4.3). Our code is available online.[3]

4.1 Data Preprocessing

For both tasks, we tested cleaning the tweet text data and combined vs. separate models per topic.

For text cleaning, we replaced common abbreviations with their full forms, like changing "I'm" to "I am" and "won't" to "will not." We then used the 'neattext' library[4] to remove URLs, emails, phone numbers, punctuation, and special characters. The text was then converted to lowercase.

In addressing the class imbalance issue, we utilized an oversampling technique. Throughout both subtasks, we inserted random minority class examples until reaching an even distribution.

4.2 Model for Argumentative Stance Classification (Subtask A)

For stance classification, we employ a BERT model for sequence classification[5] to classify the stance based on the tweet text only.

Architecture: Figure 2 shows the employed architecture. We employed the BERT tokenizer[6] for tokenizing tweets. We feed the tokens into a pretrained 12-layer BERT model for sequence classification with 12 attention heads, 110M parameters, and 768 output nodes (CLS-Token pooled from the 768 embeddings per token), with one additional linear layer and softmax-activated classification layer.

Training: The model is trained for 8 epochs on the tweets. Tested optimizers are Adam (Kingma and Ba, 2014), AdamW (Loshchilov and Hutter, 2017), and SGD (Bottou, 2010), with learning rates between $1 \cdot 10^{-5}$ and $3 \cdot 10^{-2}$.

[2] The script for downloading the dataset can be found in the shared task's Git-repository: https://github.com/ImageArg/ImageArg-Shared-Task

[3] https://github.com/webis-de/argmining23-image-arg

[4] https://github.com/Jcharis/neattext

[5] https://huggingface.co/docs/transformers/model_doc/bert

[6] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer

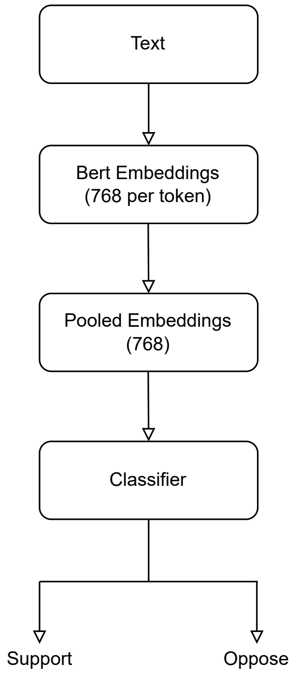


Figure 2: Our architecture for argumentative stance classification: The tweet text is tokenized, embedded through the BERT model, and then classified through a binary classification layer.

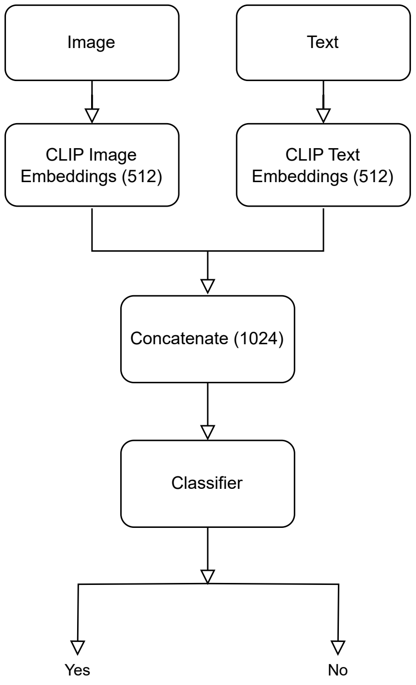


Figure 3: Our architecture for persuasiveness classification: Tweet text and image are tokenized and embedded through the CLIP model. Then features are concatenated and fed to a linear neural network, which predicts persuasiveness probability via a softmax.

Model Selection: We submitted the model with the best F1 score on the validation set, as determined by grid search, to the shared task. Namely, separate models per topic using cleaned data, the Adam optimizer with a learning rate of $3 \cdot 10^{-5}$ for the topic of "gun control", and the SGD optimizer with a learning rate of $3 \cdot 10^{-2}$ for "abortion."

4.3 Model for Image Persuasiveness Classification (Subtask B)

For image persuasiveness classification, we employ concatenated CLIP embeddings (Radford et al., 2021) of images and texts.

Architecture: Figure 3 shows the employed architecture. We used the 512-dimensional embeddings generated by CLIP for each image and text. Since CLIP can only embed texts of up to 77 word-tokens, we split longer tweets into chunks of a maximum of 77 tokens each. These chunks were then individually tokenized and stacked to a tensor to create the necessary input for CLIP's text embedding. The CLIP embeddings for text and image pairs are each represented as tensors of 512 dimensions. These embeddings are then concatenated, first the image embedding followed by the text embedding, creating a unified representation for each tweet that is 1024-dimensional. We fed the concatenated embeddings to a linear neural network, which included subsequent layers leading to a binary softmax classification layer. To investigate the influence of the features, we also tested setting all tweet texts to the empty string.

Training: The model is trained over 10 epochs for both cleaned and uncleaned tweets. We selected 10 epochs as we found that gains decreased afterward in preliminary test runs. For optimizers, we tested the same as for subtask A (Adam, AdamW, and SGD).

Model Selection: We submitted the models with the best F1 score on the validation set, determined by optimizing the learning rate and optimizer. We found that separate models per topic performed best, so we submitted those, both for cleaned and uncleaned data.

5 Results

To analyze our approach, we provide both an overview table (Table 1) and a confusion matrix (Table 2) for both subtasks.[7]

[7] Our train and dev sets have a slightly different distribution of classes compared to the original datasets, related to downloading issues.

Subtask / Model	**F1 score**		
	Abortion	**Gun control**	**Overall**
Subtask A: Argumentative Stance Classification			
Cleaned, separate *[7]	0.91	0.77	0.84
Uncleaned, separate	0.90	0.77	0.83
Cleaned, combined	0.89	0.72	0.81
Subtask B: Image Persuasiveness Classification			
Cleaned, separate *	0.56	0.54	0.56
Uncleaned, separate *	0.53	0.54	0.54
Image-only, separate	0.55	0.49	0.52

Table 1: Achieved best F1 scores for each Subtask on the Test Dataset. A "*" marks the submitted approaches.

A	**Prediction**	
Truth	Oppose	Support
Oppose	0.49	0.11
Support	0.05	0.35

B	**Prediction**	
Truth	Yes	No
Yes	0.18	0.14
No	0.17	0.50

Table 2: Confusion matrices for the best-performing models on both subtasks on the test set: Argumentative Stance (A) and Image Persuasiveness Classification (B).

5.1 Results for Argumentative Stance Classification (Subtask A)

As Table 1 indicates, our approach achieves an F1 score of 0.84,[8] highlighting its strong performance in stance classification based on tweet text. This score corresponds to the 3rd place in the competition. The confusion matrix (Table 2) shows that our model performs a bit better on supportive tweets ($0.05/0.40 \approx 0.13$ misclassification rate) than on opposing ones ($0.11/0.60 \approx 0.18$), but this might be an artifact from the specific topics.

Furthermore, we trained separate models on an uncleaned dataset and a combined model using the cleaned dataset that includes both topics. The results are displayed in Table 1. As the table shows, using separate models and cleaning the dataset results in slightly improved results.

5.2 Results for Image Persuasiveness Classification (Subtask B)

As Table 1 indicates, our approach achieves an F1 score of 0.56, reflecting mediocre performance despite winning the competition. From the confusion matrix (Table 2), we can observe that the model's performance is mixed. While it can relatively accurately identify images labelled as not enhancing the persuasiveness ($0.17/0.67 \approx 0.25$ false positive rate), it struggles to correctly identify images labelled as enhancing the persuasiveness ($0.14/0.32 \approx 0.44$ false negative rate). This discrepancy indicates that the model did hardly learn to recognize persuasive elements in the images. However, we assume that more features can improve the performance of our models, for example by identifying infographics or processing text from the images using on-screen character recognition.

Furthermore, we tested a model that did not consider the tweet text at all. As Table 1 shows, this approach performed nearly as good as our full approach (F1 score: 0.52 vs. 0.56), especially for the topic of abortion (F1 score: 0.55 vs. 0.56). As this result highlights, our model does currently barely take advantage of the actual text.

6 Conclusion

We presented the submissions of team "feeds" to the two subtasks of ImageArg 2023 (Liu et al., 2023) and results of further analyses we performed after the submission deadline.

Our approach for argumentative stance classification (subtask A) achieved a commendable F1 score of 0.84, but, as our analysis revealed, it, amongst other issues, struggled with classifying straightforward sentences like "I support gun control" or "I support abortion." Additionally, subtask A's model didn't incorporate image data. Future work could include images, for example using the VisualBERT[9] (Li et al., 2019) model, enabling classification using both text and images.

Our approach for image persuasiveness (subtask B) achieved the first position with an F1 score of 0.56. We observed that the model effectively classifies images that do not enhance persuasiveness, but struggles with identifying images that enhance the text's persuasiveness. This highlights the importance of advanced feature engineering to enhance the model's ability to identify nuanced persuasive elements within images. Moreover, we found that our classifiers perform nearly as good without considering the text at all. This emphasizes the influential role of CLIP image embeddings within the model's decision-making process. Further investigations are needed for understanding which role, if any, features from the tweet text could play in the classification of this task.

[8] Due to a mistake, we submitted predictions for only one topic by the ImageArg 2023 deadline. The values reported here are calculated using the evaluation script and data provided by the organizers after the deadline

[9] https://huggingface.co/docs/transformers/model_doc/visual_bert

Ethics Statement

We utilized the ImageArg dataset (Liu et al., 2023) without making substantial modifications to its content. The dataset was exclusively employed for participation in the ImageArg Shared Task, while adhering to the guidelines of the Twitter Developer Policy and the ACL Ethics Policy. Our primary objective was to perform stance and persuasiveness classification based on the provided text and images. Significantly, our experimental results underscore that our approach is presently unsuitable for product integration. Our primary focus remains on advancing research in this specific task.

References

Yamen Ajjour and Khalid Al-Khatib. 2021. Analysing the submissions to the same side stance classification task. In *Same Side Stance Classification Shared Task 2019*, volume 2921 of *CEUR Workshop Proceedings*.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barriere, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association (CLEF 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York. Springer.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010*, pages 177–186.

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. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, pages 2562–2571. ACM.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Erik Körner, Gerhard Heyer, and Martin Potthast. 2021. Same side stance classification using contextualized sentence embeddings. In *Same Side Stance Classification Shared Task 2019*, volume 2921 of *CEUR Workshop Proceedings*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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 (ARGMINING'22)*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.