A General Framework for Multimodal Argument Persuasiveness Classification of Tweets

Mohammad Soltani and **Julia Romberg**
Heinrich Heine University Düsseldorf
{mohammad.soltani,julia.romberg}@hhu.de

Abstract

An important property of argumentation concerns the degree of its persuasiveness, which can be influenced by various modalities. On social media platforms, individuals usually have the option of supporting their textual statements with images. The goals of the *ImageArg shared task*, held with ArgMining 2023, were therefore (A) to classify tweet stances considering both modalities and (B) to predict the influence of an image on the persuasiveness of a tweet text. In this paper, we present our proposed methodology that shows strong performance on both tasks, placing 3rd team on the leaderboard in each case with F_1 scores of 0.8273 (A) and 0.5281 (B). The framework relies on pre-trained models to extract text and image features, which are then fed into a task-specific classification model. Our experiments highlighted that the multimodal vision and language model CLIP holds a specific importance in the extraction of features, in particular for task (A).

1 Introduction

How convincing are the arguments put forward in a discussion? Are these arguments effective in persuading a dissenting voice to change its opinion or behavior? Automatically answering such questions of *argument persuasiveness* holds significant importance within the field of argument mining.

There has been a growing body of research on tasks pertaining to persuasiveness (Persing and Ng, 2015; Wachsmuth et al., 2016; Chakrabarty et al., 2019). Works like Stab and Gurevych (2014, 2017) and Habernal and Gurevych (2017) have brought persuasive essays into focus. To capture the persuasiveness of arguments based on Aristotle (2007)'s idea of logos, ethos and pathos, different annotation schemes have been developed (Duthie et al., 2016; Carlile et al., 2018; Wachsmuth et al., 2018). Moreover, phenomena of argument persuasion were examined using a variety of data sources, including online debates (Lukin et al., 2017; Durmus and Cardie, 2018; Longpre et al., 2019) and news editorials (El Baff et al., 2020).

What these works have in common is their emphasis on argumentation in textual form. However, the options for persuading the counterpart of one's own view are by no means limited to written speech (Park et al., 2014). There are further means that can be employed, usually as supplements, like images or videos (Joo et al., 2014; Huang and Kovashka, 2016; Liu et al., 2022b).

In this paper, we present our solution approach to the *ImageArg shared task* (Liu et al., 2023). We propose using a general framework to solve tasks related to argument persuasiveness in multimodal settings. The framework comprises two feature extraction modules designed for processing text and image modalities, which are subsequently inputted into a classifier. In our experiments, CLIP-extracted features (Radford et al., 2021) excelled for subtask (A), and supplementing them with additional features (ConvNeXt (Liu et al., 2022c), Reformer (Kitaev et al., 2020), ELECTRA (Clark et al., 2020), LayoutLM (Xu et al., 2020), CamemBERT (Martin et al., 2020), Swin V2 (Liu et al., 2022a)) proved most beneficial for subtask (B).

We begin with a brief description of task and dataset (§2), followed by a detailed description of our methodology (§3). We then present the experimental results (§4) and analyze the errors that occur (§5). In addition, we report progress on our approach in the post-evaluation phase, which has enabled us to further improve classification performance (§6). Finally, we draw a conclusion and make recommendations for future work (§7).

2 Task Description

The shared task relies on ImageArg (Liu et al., 2022b), a multimodal dataset for argument persuasiveness. It consists of English-language argumentative tweets supported by images as provided by users. The version of the dataset used for the shared

Proceedings of the 10th Workshop on Argument Mining, pages 148–156

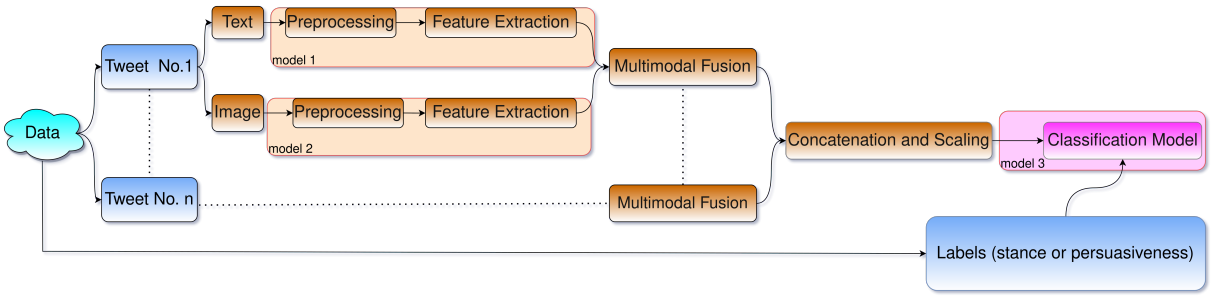

Figure 1: Framework design: model 1 and 2 extract the text and image features for each tweet as vectors of sizes a and b. Multimodal fusion combines these into a single vector of size c, with $c = a + b$. The n tweet feature vectors then jointly form a matrix $C \in \mathbb{R}^{c \times n}$. Along with the n task-specific labels, they serve as input for model 3.

task includes two subtasks: *Argumentative Stance (AS) Classification (Subtask A):* Given a tweet and an accompanying image, predict the stance (either *support* or *oppose*) that the tweet takes on a particular topic. *Image Persuasiveness (IP) Classification (Subtask B):* Given a tweet and an accompanying image, predict whether or not the image makes the tweet more persuasive (either *yes* or *no*).

Table 1 gives an overview of the dataset[1], which covers two controversial topics, *abortion* and *gun control*. Evidently, there is an imbalance in the data pertaining to both subtasks. For AS, while both stances reach a balance on gun control, opposition clearly prevails on abortion. As for IP, adding images only contributes to tweet persuasiveness in about one-third of the cases for both topics.

		AS			IP	
		abortion	gun control		abortion	gun control
train	total	887	914	total	887	914
	supp.	243	471	yes	278	250
	opp.	644	443	no	609	664
dev	total	100	96	total	100	96
	supp.	19	51	yes	26	31
	opp.	81	45	no	74	65
test	total	150	150	total	150	150
	supp.	33	85	yes	53	53
	opp.	117	65	no	97	97

Table 1: Overview of the data distribution among the two topics and for the different data splits.

3 Methodology

Motivated by Liu et al. (2022b), we developed a versatile framework (illustrated in Figure 1) that takes tweet texts and images as input, extracts features for both modalities, and feeds the combined features into a classification model. This framework is designed to work readily for both tasks and comprises the following stages:

3.1 Multimodal Feature Extraction & Fusion

The multimodal feature extraction consists of three steps that are iterated for every tweet in the dataset.

Feature Extraction from Text Each tweet text is first tokenized. Using some pre-trained language model (*model 1*), text features are then extracted in order to represent the semantic information.

Feature Extraction from Image In parallel, each tweet image is readied for feature extraction through transformation, resizing, normalizing, and adjusting dimensions. Subsequently, the prepared image is processed by a specified pre-trained model (*model 2*) to extract image features.

Early Multimodal Fusion We then combine features from both modalities by concatenating them along the last dimension according to the early fusion strategy suggested by Boulahia et al. (2021) for creating a unified representation that combines image and text information.

3.2 Feature Concatenation and Scaling

We retain combined features of all data instances in an array and enhance their impact during learning by scaling them (Singh and Singh, 2020). For this, we re-scale each feature by its maximum absolute value, keeping them in a range between -1 to 1.

3.3 Classification

In a final step, the tweet representation obtained by the previous process serves as input to a classification model (*model 3*). This model is trained using the given training data and the corresponding labels for the respective task (either AS or IP).

[1]Our statistics differ slightly from the organizer's data due to inconsistencies in the downloading process.

Model Type	Model Architectures
Text (model 1)	Sentence-BERT, BERT, RoBERTa, ALBERT, DistilBERT, ELECTRA, XLNet, CTRL, Longformer, DeBERTa, XLM-RoBERTa, FlauBERT, DialoGPT, LayoutLM, Funnel-Transformer, MBart, CamemBERT, Reformer, Transformer-XL, GPT3, CLIP, ALIGN
Image (model 2)	AlexNet, ConvNeXt, DenseNet, EfficientNet, EfficientNetV2, GoogLeNet, Inception v3, MaxViT, MnasNet, MobileNetV2, VGG, MobileNetV3, RegNet, ResNet, ResNeXt, ShuffleNet v2, SqueezeNet, Swin Transformer, ViT, Wide ResNet, CLIP, ALIGN
Classifier (model 3)	Logistic Regression, XGBoost, Gradient Boosting, AdaBoost, CatBoost, LightGBM, MLPClassifier, SGDClassifier, SVM (with kernels: linear, poly, rbf, sigmoid), Gaussian Naive Bayes, EasyEnsemble, KNeighborsClassifier, Random Forest, Decision Trees, Extra Trees, RUSBoostClassifier, BalancedBaggingClassifier, BalancedRandomForestClassifier, PassiveAggressiveClassifier, GaussianProcessClassifier with kernel RBF, RidgeClassifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis

Table 2: Summary of the models utilized in our experiments.

		abortion			gun control					
	attempt	model 1	model 2	model 3	model 1	model 2	model 3	train mode	F_1 (dev)	F_1 (test)
AS	1	CLIP32	CLIP32	AdaBoostClassifier	CLIP32	CLIP32	AdaBoostClassifier	separate	0.9254	0.8142
	2	CLIP32	CLIP32	AdaBoostClassifier	CLIP32	CLIP32	XGboost+GradientBoosting	separate	**0.9333**	**0.8273**
	3	CLIP32	CLIP32	AdaBoostClassifier	CLIP32	CLIP32	RUSBoostClassifier	separate	**0.9333**	0.8000
	4	CLIP32	CLIP32	XGboost+GradientBoosting	CLIP32	CLIP32	XGboost+GradientBoosting	joint	0.9142	0.8093
	5	CLIP32	CLIP32	SVM-Poly	CLIP32	CLIP32	SVM-Poly	joint	0.9197	0.7782
IP	1	CLIP32	CLIP32	SVM-Poly	CLIP32	CLIP32	SVM-Poly	joint	0.6605	0.4875
	2	CLIP32	CLIP32	SGD	CLIP32	CLIP32	SGD	separate	0.6552	0.4545
	3	CLIP_L_14	CLIP_L_14	SVM-Poly	CLIP_L_14	CLIP_L_14	SVM-Poly	joint	0.6721	0.4762
	4	CLIP32	CLIP32	SGD	Convnext_small	REL	LogisticRegression	separate	**0.6726**	0.4778
	5	CLIP32	CLIP32	SVM-Poly	Convnext_small	REL	LogisticRegression	separate	0.6667	**0.5281**

Table 3: Selected submissions and their performance on dev and test for both tasks. Participants were free to decide whether they wanted to create a cross-topic model (train mode: joint) or topic-specific ones (train mode: separate).

4 Experiments

4.1 Model Selection for Submission

We conducted extensive experiments using our framework with a variety of pre-trained models from both PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) libraries. Our Python implementation is available at https://github.com/mohsoltani/GFMAP.

In fact, we examined more than 300 different combinations of these models for each subtask and topic. For our classification approach, we experimented either with a single classifier or with ensemble learning (Dong et al., 2020), combining two or more classifiers. Table 2 provides an overview of the different models we tested.

Using CLIP as a text and image model, we conducted experiments with all of the listed classification models. Subsequently, we investigated the performance of the top classification models for other combinations of pre-trained models, where Logistic Regression was found to be the most effective classification model. The best hyperparameters of the classification model were determined by trial and error (an overview is provided in Appendix A).

Among these experiments, we identified the best-performing models, which were then candidates for further experimentation involving the joint consideration of topics within each subtask. Ultimately, our submissions for the shared task at hand consisted of the top five performing models derived from our thorough experimentation.

4.2 Results

Table 3 shows our five submissions to both tasks. In AS, attempt 2 performed best, using CLIP[2] to extract the features that are subsequently fed into the classifier (AdaBoost for abortion, an ensemble of XGBoost and GradientBoosting for gun control). While our most effective strategy utilizes models tailored to specific topics, attempt 4 demonstrates that a generalized model is only slightly inferior to customized solutions (0.8273 vs. 0.8093 F_1).

The best approach for IP shows that in this case the choice of feature extraction models is different for the topics. While CLIP is again suitable for abortion, a combination of ConvNeXt[3] and REL (a concatenation of features extracted through Reformer[4], ELECTRA[5] and LayoutLM[6]) is the best choice for gun control, leading to an F_1 score of 0.5281. Cross-topic models score significantly lower on this task, which may indicate that the role of imagery in making textual arguments more persuading is topic-dependent.

5 Error Analysis & Discussion

In the following, we analyze the outputs of our best model for AS and IP in terms of misclassifications:

[2] CLIP32: https://huggingface.co/openai/clip-vit-base-patch32; CLIP_L_14: https://huggingface.co/sentence-transformers/clip-ViT-L-14

[3] Convnext_small: https://pytorch.org/vision/stable/models/generated/torchvision.models.convnext_small.html

[4] https://huggingface.co/google/reformer-crime-and-punishment

[5] https://huggingface.co/google/electra-small-discriminator

[6] https://huggingface.co/microsoft/layoutlm-large-uncased

5.1 Argumentative Stance Classification

The main reasons behind the most prevalent mistakes are:

Sarcasm, Humor, & Lack of Information In some cases, our approach faces difficulties in discerning a tweeter's true intent. One reason for this is sarcastic tweets: If a tweet seems to express positivity, but the tweeter takes the opposite view, misclassifications occur. Likewise, very short tweets tend to be misclassified, especially when negative words dominate but the overall stance is support. However, as we have noted, indirect communication (i.e., the speaker does not explicitly express his or her intentions or feelings) using humor or sarcasm can confuse not only the model but also the human audience when it comes to understanding the author (Appendix B.1 gives further insights).

Specifics of the Gun Control Topic In the area of gun control, there are two opposing groups and one supporter group: (1) The first group of critics advocates for a world without guns. (2) The second group of critics champions personal freedom and opposes any restrictions on the sale or use of guns. (3) The supporters advocate for regulated sales and usage of firearms.

In fact, the interesting dynamic surrounding gun control is that both groups of detractors are opposed to the proponents but also hold conflicting views among themselves. This complexity can make it challenging to discern the intention behind certain words or phrases in a tweet, such as "end of gun violence". Depending on the context and tweeter's specific stance, this phrase could potentially be interpreted in two different ways: It could be seen as a call for regulations and controls on the sale and use of guns to put an end to gun violence. This interpretation aligns with the stance of supporting gun control. Alternatively, it could be interpreted as a call for complete prohibition of selling and using guns, with the aim of eliminating gun violence entirely. This interpretation aligns with the stance of opposing guns altogether. An even more complicated scenario arises when considering this phrase in the context of using firearms for defense purposes: There seems to be a shared belief among groups 2 and 3 that the presence of a firearm may occasionally reduce the likelihood of firearm-related violence when used for defense.

Facing such scenarios, it is difficult to decide definitively whether we should take the supportive group stance, since a statement may not have direct relevance to an opposing group. In certain cases, discerning between the supportive group and one of the opposing groups can be quite challenging. We are dealing with a triangular arrangement of groups that must be classified into two classes. For further insights, please refer to Appendix B.2.

Ambiguities in Labels In certain instances where there are deviations between the predicted and gold labels, we found it difficult to confirm ourselves that the predicted stance is definitely incorrect (see Appendix B.3 for more details).

5.2 Image Persuasiveness Classification

Assessing whether an image enhances the tweet's persuasiveness presents a significant challenge – even for humans. The methods for visually representing or amplifying the stance of a tweet offer a variety of options compared to pure text:

Text Within Image A common approach is to insert a repetition of the tweet text or other relevant text in the image. This also allows the text's impact to be enhanced through visual effects such as image transformations, shading, different letter styles, adding text borders, colors, and background changes. We found that our best model developed a tendency to classify images showing only text as persuasive. However, the gold standard also contains many cases where this type of image was coded as not contributing to persuasiveness. We suspect that our model's behaviour is due to the fact that it is not able to extract and understand text from images. Therefore, in these cases, the model cannot make decisions based on linguistic semantics, but only on the structure of the image.

Image Persuasion Strategies Further strategies involve illustrating cases, consequences, or outcomes related to the text argument. A more intricate approach we found in the analysis visualizes counterexamples for opposing points of view.

It is difficult to objectively determine whether these methods are compelling or not, as images provide extensive creative freedom, allowing words and phrases to take on different visual forms. In addition, image effects (see e.g. Szeliski, 2022) such as occlusion, distinct object placements, viewpoint variations, deformations, background clutter, exposure bracketing, and morphing can change the illustrating form, consequently influencing the viewer's perception in various ways. Given the multitude of

phenomena, we abstain from delving into specifics within the scope of this article.

Human Label Variation Human perceptions are often subjective and influenced by emotions, personal preferences or cultural backgrounds (Pettersson, 1982). For example, depictions of scenes such as protests can evoke different reactions depending on cultural norms and personal experiences. While protests are welcome in some cultures, they can be prohibited in others, resulting in either excitement or a sense of normality. Annotators with different thinking styles, such as holistic and analytical (Li et al., 2022), may also make different judgments when considering the background context or focusing solely on the objects in an image (for examples see Appendix B.4). Liu et al. (2022b) annotated image persuasiveness by assigning an aggregate label to establish a unified scoring. To account for different valid perceptions of persuasiveness resulting from the previously listed reasons, this approach may be insufficient and deserves reconsideration.

Impact of Image In the process of constructing the dataset, when a tweet's text was rated as extremely persuasive, the supplementary persuasiveness attributed to attached images was devalued, eventually resulting in a *no* label. This may lead the machine learning approach astray since the image itself can be highly persuasive in its own right.

6 Additional Experiments

As can be seen from the results presented so far, predicting IP in particular presents a challenge. For this reason, we present additional experiments that we conducted as a follow-up to the shared task.

In our experimental efforts, we obtained notably positive results when using CamemBERT as text model, particularly for abortion in combination with ConvNeXt or Swin Transformers V2 as image model. Given the significant disparity between our dev and test scores for IP in the shared task submissions, we proceeded to conduct additional experiments with various adaptations of this model in order to find more robust models.

It turned out, that employing camembert-base[7] to extract text features and swin_v2_s[8] to extract image features for the abortion topic, while retaining the proven combination of REL and ConvNeXt for the gun control topic, resulted in promising results. The classifier was Logistic Regression. With this setup, we managed to attain an F_1 score of 0.5941 for the test set, while the F_1 score for the dev set was 0.5950. As can be seen, the approach significantly increases previous test scores (cf. Table 3) while obtaining robust results across dev and test set. Our finding suggests that model performance should generalize to further in-domain datasets.

We performed further experiments, eventually achieving test scores above 0.66. At the same time, however, the dev performance deviated strongly downwards in these cases. Despite the very encouraging results, additional investigations are needed in order to ensure reliable performance.

7 Conclusion & Future Work

On social media, users have the freedom to use informal, formal, or mixed styles of language, and to incorporate elements such as hashtags, mentions, links to websites, and emojis. In addition, images can be used to substantiate textual statements. This variety presents a challenge when trying to classify argument stances and their persuasiveness from sources such as X (formerly known as Twitter). As the analysis of our approach was able to reveal, the prevalence of sarcasm and the limited information content in tweets substantially complicates the classification. This observation underscores the need for further improvement of models tailored to the specific characteristics of social media data.

In the context of classifying the additional persuasive power of images over text, it is crucial to use models that not only extract image features or detect objects in images, but can also extract the attitude and persuasion expressed through the images themselves. This necessitates the design of visual argument extraction models. The particular difficulty of evaluating the argumentative persuasiveness of images, as well as the inherently subjective nature of the task, require special attention.

What is more, due to the training dataset's limited size, it becomes challenging to differentiate learned image features at a granular level from those in other images. A larger dataset may assist us in improving classification results, particularly to overcome the challenges outlined in Section 5.

Possible research directions also include delving into the applicability of CamemBERT to English texts and exploring the reasons why this model surpasses English models in the task at hand.

[7] https://huggingface.co/camembert/camembert-base

[8] https://pytorch.org/vision/main/models/generated/torchvision.models.swin_v2_s

Acknowledgements

Julia Romberg is funded by the Federal Ministry of Education and Research of Germany, project CIMT/Partizipationsnutzen of the funding priority Social-Ecological Research (funding no. 01UU1904). Responsibility for the content of this publication lies with the authors.

References

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse*. Clarendon Aristotle seris. Oxford University Press. (George A. Kennedy, Translator).

Said Boulahia, Abdenour Amamra, Mohamed Madi, and Said Daikh. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32:121.

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. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*. OpenReview.net.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14:pages 241–258.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045. Association for Computational Linguistics.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of the Sixth International Conference on Computational Models of Argument (COMMA 2016)*, pages 299–310. IOS Press.

Roxanne El Baff, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Persuasiveness of news editorials depending on ideology and personality. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 29–40. Association for Computational Linguistics.

Steven Greene, Melissa Deckman, Laurel Elder, and Mary-Kate Lizotte. 2022. Do moms demand action on guns? Parenthood and gun policy attitudes. *Journal of Elections, Public Opinion and Parties*, 32(3):655–673.

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

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 (CVPR) Workshops*, pages 73–79. IEEE.

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 (CVPR)*, pages 216–223. IEEE.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*. OpenReview.net.

Hao Li, Ting Wang, Yi Cao, Lili Song, Youbo Hou, and Yizhi Wang. 2022. Culture, thinking styles and investment decision. *Psychological Reports*, 125(3):1528–1555.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022a. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009. IEEE.

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*. Association for Computational Linguistics.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022b. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18. International Conference on Computational Linguistics.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022c. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976,. IEEE.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219. Association for Computational Linguistics.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552. Association for Computational Linguistics.

Rune Pettersson. 1982. Cultural differences in the perception of image and color in pictures. *Educational Technology Research and Development*, 30(1):43–53.

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*, pages 8748–8763. PMLR.

Dalwinder Singh and Birmohan Singh. 2020. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524.

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. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Richard Szeliski. 2022. *Computer Vision: Algorithms and Applications*. Springer Nature Switzerland AG.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

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*, pages 1192–1200. Association for Computing Machinery.

A Hyperparameter Fine-tuning

Table 4 outlines the hyperparameters used in the classification models of our submissions.

B Error Analysis: Details

B.1 Sarcasm

We have noticed that some tweets can be infused with sarcasm, such as: *Gov. Ralph 'Coonman' Northam proud to sign a slew of new 'commonsense gun safety measures' that will save lives*

		abortion		gun control	
	attempt	classifier(s)	parameters	classifier(s)	parameters
AS	1	AdaBoostClassifier	base_estimator= DecisionTreeClassifier(max_depth=2), n_estimators=150, learning_rate=0.2, algorithm='SAMME'	AdaBoostClassifier	base_estimator= DecisionTreeClassifier(max_depth=2), n_estimators=150, learning_rate=0.3, algorithm='SAMME'
	2	AdaBoostClassifier	base_estimator= DecisionTreeClassifier(max_depth=2), n_estimators=150, learning_rate=0.2, algorithm='SAMME'	XGboost+GradientBoosting	XGB : max_depth=3, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, reg_lambda=0.1 GradientBoosting: learning_rate=0.2, n_estimators=80, random_state=42 Voting:'xgb', 'gb', voting='soft', weights=[3, 1]
	3	AdaBoostClassifier	base_estimator= DecisionTreeClassifier(max_depth=2), n_estimators=150, learning_rate=0.2, algorithm='SAMME'	RUSBoostClassifier	n_estimators=150, random_state=42, learning_rate=0.18, sampling_strategy='not majority'
	4	XGboost+GradientBoosting	XGB : max_depth=2, learning_rate=0.3, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, reg_lambda=0.12 GradientBoosting: learning_rate=0.4, n_estimators=80, random_state=42 Voting:'xgb', 'gb', voting='soft', weights=[5, 2]	XGboost+GradientBoosting	XGB : max_depth=2, learning_rate=0.3, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.1, reg_lambda=0.12 GradientBoosting: learning_rate=0.4, n_estimators=80, random_state=42 Voting:'xgb', 'gb', voting='soft', weights=[5, 2]
	5	SVM-Poly	kernel='poly', degree=2, coef0=0.6	SVM-Poly	kernel='poly', degree=2, coef0=0.6
IP	1	SVM-Poly	kernel='poly', degree=2, coef0=0.02, shrinking=False, probability=True	SVM-Poly	kernel='poly', degree=2, coef0=0.02, shrinking=False, probability=True
	2	SGD	alpha=0.0344, random_state=42	SGD	alpha=0.05, random_state=42
	3	SVM-Poly	kernel='poly', random_state=42, coef0=0.17	SVM-Poly	kernel='poly', random_state=42, coef0=0.17
	4	SGD	alpha=0.0344, random_state=42	LogisticRegression	by default
	5	SVM-Poly	kernel='poly', degree=2, coef0=0.25	LogisticRegression	by default

Table 4: Hyperparameters employed for tuning the classification models in the Python implementation.

https://t.co/3toCAPRO1b *via @twitchyteam*[9]. In the test set, this tweet has been labeled as opposing gun control. However, there is no clear evidence that the tweet explicitly expresses a contrary view, as the presence of sarcasm might be a factor to consider in this case given the particular use of quotation marks.

B.2 Specifics of the Gun Control Topic: A Triangular Perspective

The three groups in this triangular arrangement advocate for a gun policy characterized by: (1) absence of guns (opposing gun control), (2) unrestricted use of guns (opposing gun control), and (3) regulated and legally permissible use of guns (supporting gun control). Table 5 shows differences and similarities in opinion among all three groups.

Groups	**Similarities in Opinion**
1-2	Oppose to regulation of usage and selling guns
1-3	Safety Measures to Protect Lives from Gun Violence
2-3	Existence of guns
Groups	**Differences in Opinion**
1-2	Existence of guns
1-3	Existence of guns
2-3	Stringent Regulations for Gun Control

Table 5: Comparison of the three groups (gun control).

In certain cases, it is not straightforward to associate a sentence or tweet to one of these groups. To illustrate these challenges, we analyze the following tweet from the perspectives of all three groups: *Women are five times more likely to be killed by their abuser if there is a gun present. We can prevent tragedy. We can work together and help people. We need #gunsenselegislators. We need @JoeBiden and @KamalaHarris. #VAWA #DisarmHate #ERPO #OneThingToDo #expectUs @MomsDemand*[10].

Challenges arise in the first sentence: *Women are five times more likely to be killed by their abuser if there is a gun present.* This can be assigned to the first group that fights for the absence of guns. However, it is also conceivable that the argument could be used by the other groups. Group 3, supporters of gun control, accept the existence of guns but argue that without strict laws, the presence of guns can lead to such violence. Group 2, which criticizes gun regulation, may argue that such regulations could create situations in which women, by taking advantage of the law, might provoke abusers to use violence against them. They seem to hold the view that restrictions on gun control can lead to acts of violence.

To determine the true stance of the tweet, we analyze the following sentences. The next two sentences can align with each perspective, supporting their respective stances. The essential sentence in this tweet is as follow: *We need #gunsenselegislators.* This statement can be linked to supportive groups, as the term "gunsense" refers to individuals advocating for gun control. In addition, the tweeter

[9]1249087853558222850: https://t.co/2KiRh4RAEA

[10]1296267688310906880: https://t.co/ydP75LeEmQ

mentioned @MomsDemand, which reinforces the same notion (Greene et al., 2022). They are actively involved in promoting stricter gun control regulations and reducing gun violence.

This example demonstrates that the presence of a specific word or phrase can be decisive in indicating the actual stance of a tweet, even when other sentences could be associated with other or all groups. This complexity poses a major challenge for argumentation mining models.

B.3 Ambiguities in Labels

In the subsequent cases, comprehending the motivations behind the assigned stances in the test set proves to be a challenging task:

Abortion Our model has classified the following tweet as supportive, whereas in the gold standard it is labeled as opposing abortion. A closer look reveals, that it is a promotional tweet promoting clinic's abortion pills: *Abortion pills are effective, and you could have your abortion in Bethal with pills anytime at an affordable price. Contact +27727793390. https://t.co/ffj25TRLIB0*[11]. This tweet emphasizes women's right to make decisions about their own bodies and, thus, seems to be in line with the positions of groups promoting abortion rights.

The following tweet criticizes the dismantling of abortion rights but has been labeled as opposing abortion, while our model predicts it as supporting abortion: *Overturning Roe v. Wade will not reduce abortions but become a contributing factor in increasing poverty, dismantling Civil Rights, and literally moving the country back decades. @mskathykhang #SCOTUS Sign the #PledgetoPause: https://t.co/Dtf8a6SSSR*[12].

Gun Control Another example of a possible misinterpretation of a tweet in the test set is: *Women are five times more likely to be killed by their abuser if there is a gun present. We can prevent tragedy. We can work together and help people. We need #gunsenselegislators. We need @JoeBiden and @KamalaHarris. #VAWA #DisarmHate #ERPO #OneThingToDo #expectUs @MomsDemand*[13]. While the gold label is oppose, the phrases "gunsenselegislators" and "MomsDemand" refer to actions advocating gun control measures. Our model has classified the aforementioned tweet as supportive of gun control.

Irrelevance to Topic *Vaccines save. Stupidity kills. #antimask #antimaskers #karensgonewild #karenmemes #trump2020 #vaccines #election2020 #prochoice #bidenharris2020 #memes #racism #covid19 #endracism #prolife #wearadamnmask #hoax #trumpvirus*[14]. This tweet refers to the topic of COVID vaccination. Although the tweet is labeled as supporting abortion in the test set (and our model predicted it as opposing abortion), there is no clear indication in the tweet to express support or opposition to abortion.

B.4 Challenging Examples in Image Persuasiveness

As noted in the discussion in subsection 5.2, a broader range of methods are available to convey the attitude of a tweet through images compared to text alone. In the test set, following tweets labeled as not persuasive were predicted as persuasive by our best model:

Abortion: *New year. New opportunities to end abortion. Are you with us? RT if you stand with preborn children. #EndAbortion #ProLife*[15]. The corresponding image mirrors the message "New Year. New opportunities to end abortion" underpinned with the illustration of a smiling pregnant woman to enhance persuasiveness (in our subjective perception).

Gun Control: *Gun stores are not essential businesses during the #COVID19 crisis. Arming the medical community with the equipment they need is. Sign this petition urging The Trump Admin to remove gun stores from that list.*[16]. The corresponding image shows a woman wearing a red shirt with a "MomsDemand" symbol to encourage signing. Again, this can be perceived to strengthen the urge for a petition to remove gun stores from the list of essential businesses during the pandemic.

[11] 1331187788096606208: https://t.co/ZFoAGRje4T
[12] 1022572268147208192: https://t.co/TS6ZNBbR8v
[13] 1296267688310906880: https://t.co/ydP75LeEmQ
[14] 1335685471205289989: https://t.co/UfJ74ayA9S
[15] 1347211895674122245: https://t.co/os3O4lwPa2
[16] 1245045552984674304: https://t.co/05vcbnrH6r