Argumentative Stance Prediction: An Exploratory Study on Multimodality and Few-Shot Learning

Arushi Sharma*, **Abhibha Gupta***, **Maneesh Bilalpur***
School of Computing and Information
University of Pittsburgh
{arushi.sharma, abg96, mab623}@pitt.edu

Abstract

To advance argumentative stance prediction as a multimodal problem, the *First Shared Task in Multimodal Argument Mining* hosted stance prediction in crucial social topics of gun control and abortion. Our exploratory study attempts to evaluate the necessity of images for stance prediction in tweets and compare out-of-the-box text-based large-language models (LLM) in few-shot settings against fine-tuned unimodal and multimodal models. Our work suggests an ensemble of fine-tuned text-based language models (0.817 F1-score) outperforms both the multimodal (0.677 F1-score) and text-based few-shot prediction using a recent state-of-the-art LLM (0.550 F1-score). In addition to the differences in performance, our findings suggest that the multimodal models tend to perform better when image content is summarized as natural language over their native pixel structure and, using in-context examples improves few-shot performance of LLMs.

1 Introduction

Argumentative stance studies related to ideological topics offer valuable insights into complex dynamics of opinion, belief and discourse in various domains. These insights have far-reaching implications, extending their influence over areas including public opinion, social dynamics, and policy efficacy. By predicting the stance in real-time, policymakers and stakeholders can get immediate feedback on public reaction to new proposals or laws, allowing them to make timely and informed decisions.

Argumentative stance prediction is becoming a major endeavor in multiple research fields as the reliance on sentiment detection may be sub-optimal (Reveilhac and Schneider, 2023). While the stance prediction task appears similar to sentiment analysis, it has many theoretical differences. Sentimental analysis primarily focuses on emotions, whereas the stance prediction need not necessarily coincide with the sentiment directed towards the target. Stance prediction for sensitive and polarizing topics can be more challenging, particularly within the brief context of informal social media text (Alturayeif et al., 2023).

Previous studies have primarily concentrated on examining stance prediction in textual modalities (Alturayeif et al., 2023; Hosseinia et al., 2020). However, an increasing number of recent works are widening the focus to include other modalities, such as images. Since multimodality helps us understand language from the modalities of text, vision and acoustic, (Zadeh et al., 2018), the application of multimodal inputs in argumentative stance prediction seems promising.

Towards the perpetuation of multimodality in argumentative stance prediction as a part of the ImgArg 2023 (Liu et al., 2023) challenge, we explore the following questions using a dataset of tweets on gun control and abortion topics:

1. How well does language as a stand-alone modality perform at argumentative stance prediction?

2. Does incorporating image information improve prediction performance?

3. How do Large-Language Models (LLMs) in few-shot setting compare against fine-tuned unimodal and multimodal models?

Our work shows that an ensemble of fine-tuned language models performs the best for argumentative stance prediction from tweets. Incorporating image information into text using state-of-the-art multimodal models does not outperform the ensemble model. LLMs (particularly, LLaMA-2) in few-shot setting exhibit high recall but suffer from low precision. Though using in-context examples

* Equal contribution

Proceedings of the 10th Workshop on Argument Mining, pages 167–174

in few-shot setting improves performance, they underperform the ensemble model.

2 Related Work

Existing work has explored the interplay between stance and sentiment to enhance stance detection. (Sobhani, 2017) investigated the relationship between stance and sentiment, utilizing SVM with N-gram, word embedding, and sentiment lexicon features. They concluded that while sentiment features offer utility, they are insufficient on their own for effective stance detection. Meanwhile, (Hosseinia et al., 2020) showcased the prowess of bidirectional transformers in achieving competitive performance without fine-tuning, harnessing sentiment and emotion lexicons. Their findings show the efficacy of sentiment information, as opposed to emotion, in discerning the stance.

(Alturayeif et al., 2023) conducted an extensive analysis of 96 primary studies spanning eight machine learning techniques for stance detection and its applications. The analysis suggests that deep learning models with self-attention mechanisms were found to be frequently outperforming the traditional machine learning models such as SVM, and emerging techniques like few-shot learning and multitask learning were increasingly applied for stance detection.

Multimodal stance detection is being increasingly used for social applications such as rumor verification (Zhang et al., 2021) and identifying public attitudes towards climate change on Twitter (Upadhyaya et al., 2023). Despite recent advancements in multimodal language models (Wang et al., 2023), the use of image modality for stance detection remains an underexplored area. Our work conducts an exploratory study to investigate the necessity of multimodal models for stance detection and compares different ways to incorporate image information into text modality.

3 Dataset and Task

The ImgArg dataset (Liu et al., 2022) is a part of the *Multimodal Argument Mining* (Liu et al., 2023) competition. Curated with the goal of expanding argumentation mining into multimodal realm, the dataset consists of Twitter texts along with their images from two topics–gun control and abortion. Each text-image pair corresponding to a tweet are annotated with a stance (support or oppose) along with its persuasiveness (no persuasiveness to extremely persuasive). In this paper, we focus on the stance prediction task. Briefly, the task can be described as given an image-text pair corresponding to a tweet, predict if it supports or opposes the topic.

It is important to note that while the gun control dataset is balanced, the abortion dataset is imbalanced by a 1:3 support:oppose stance ratio. The gun control and abortion training sets are 920[1] and 891 tweets respectively. Both datasets have an equal number of tweets in the validation (100 tweets) and test (150 tweets) sets.

4 Approach

To predict argumentative stance over multimodal tweets from gun control and abortion topics, we leverage three different ideas. We explore an ensemble of LLMs against its constituent models, incorporate image information through multimodal models as well as evaluate out-of-the-box LLMs in few-shot setting. This section describes the experimental approaches used in the process. Further details can be found in the appendix.

4.1 Ensemble Stance Prediction

Individual language models have demonstrated their superior performance across a variety of tasks. However, ensemble methods tend to perform better (Jiang et al., 2023) than their constituent models. To explore this idea, we evaluated text-based language models such as XLNet (Yang et al., 2019), XLM-RoBERTa (Conneau et al., 2019), Transformer XL (Dai et al., 2019), DeBERTa-v2 (He et al., 2020), BLOOM-560M (Scao et al., 2022). Since the dataset is a collection of tweets, conventional problems such as very long sequence length were non-existent.

Ensemble decisions were based on the weighted sum of constituent model predictions. Each model prediction was weighted by its F1-score on the validation set in order to assign a higher weight to the model that performed better on the validation set. This weighted sum is then thresholded by the F1-score averaged across models for final prediction. In our study, XLNet and BLOOM-560M received the predominant weights for attaining the highest F1 score on abortion and gun-control datasets respectively.

[1]The organizers reported 923 tweets, however, three tweets were dropped because of download issues.

4.2 Multimodal Stance Prediction

To evaluate the utility of image augmentation to text and the possible ways to achieve this, we studied models from different frameworks. The ViLT (Kim et al., 2021) is a popular vision-language transformer model with reduced computational overhead because of its convolution-free architecture. FLAVA (Singh et al., 2022), a multimodal model built to generalize to both vision tasks and language tasks. Both models were fine-tuned over the gun control and abortion datasets for the support stance prediction task.

Recent vision-language pre-trained models such as instructBLIP (Dai et al., 2023) have demonstrated solving image-centric tasks through natural language. We leverage this instruction-based summarization of image content with instructBLIP. Specifically, we summarize each image using the *briefly describe the content of the image* instruction. The resulting textual descriptions of images along with their corresponding tweets were used for stance prediction by fine-tuning a RoBERTa (Liu et al., 2019) classifier followed by early fusion. We refer to this configuration (Figure 1) as the multimodal RoBERTa.

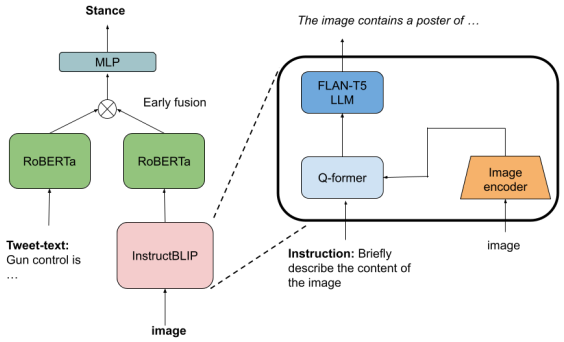

Figure 1: Multimodal RoBERTa configuration. The figure shows the input image summarized as text through instructBLIP and then used to fine-tune the RoBERTa model together with the tweet-text. Shared color between RoBERTa models indicates tied weights.

4.3 Few-Shot Stance Prediction using LLMs

Few-shot prediction typically involves using relevant examples during training to learn a new concept that was not included in pretraining. It has been a success not just in conventional language-based tasks but also in multimodal tasks (Luo et al., 2020). The large and diverse pre-training corpora used in training the foundation models is attributed as one of the reasons for their success in learning with a limited resources paradigm. Using LLaMA-2 (Touvron et al., 2023), we performed stance prediction in few-shot setting. LLaMA-2 was chosen because of its open-source implementation that outperforms commercial large-scale GPT-3 (Brown et al., 2020) with fast inference.

Choice of few-shot examples: Arguments can be made from different viewpoints or themes. For example, gun control can be referred to from ordinary themes such as the constitutionally granted *right to bear arms, governmental overreach* to targeted themes or experiences such as *school shootings*. We believe that the ImgArg dataset encompasses these diverse themes and wish to leverage the in-context examples that correspond to the same theme for few-shot experiments. We identify the themes in the training set using k-means clustering and pick examples from the same theme cluster during inference. Performance on the validation set was used as a benchmark to identify 12 clusters for the gun control dataset and 13 clusters for the abortion dataset. The manually identified cluster themes are presented as Table 8 in the Appendix A.

4.4 Experimental Setup

The imbalance in the abortion dataset is addressed using a weighted cross-entropy loss. Increased weightage was allocated to the minority category loss. The models were trained using HuggingFace (Wolf et al., 2020) on two A100 NVIDIA GPU environment[2]. Hyperparameters (learning rate, scheduler and weight decay) were optimized for the validation set and performance is reported as precision, recall and F1-score for the support stance and oppose on the test set. More experimental details are shown in Appendix A section.

5 Results

5.1 Support Stance

Table 1 compares the support class performance of individual language models against their ensemble model. The ensemble model used BLOOM-560M as it performed better than its larger counterpart on the validation set. The constituent models typically have a better recall but low precision, the ensemble model improves precision with a limited drop in the recall. Best performance was observed with the

[2]The code is available at: https://github.com/arushi-08/EMNLP-ImageArgTask-PittPixelPersuaders

Model	Precision	Recall	F1
XLNet	0.619	0.924	0.741
BLOOM-1B	0.760	0.660	0.710
BLOOM-560M	0.707	0.898	0.791
Transformer-XL	0.571	0.881	0.693
DeBERTa-v2	0.560	0.710	0.630
XLM-RoBERTa	0.650	0.880	0.750
Ensemble	**0.743**	**0.906**	**0.817**

Table 1: *Support* stance performance using text-based transformer models.

ensemble of unimodal language models with 0.817 F1-score.

Model	Precision	Recall	F1
ViLT	0.680	0.432	0.528
FLAVA	0.570	0.650	0.610
Multimodal RoBERTa	**0.531**	**0.932**	**0.677**

Table 2: *Support* stance performance using image-text multimodal transformer models.

Multimodal RoBERTa and FLAVA sacrificed precision for recall (shown in Table 2) upon fine-tuning. Both multimodal RoBERTa and FLAVA that leverage images in pixel-space achieve a recall of 0.932 and 0.650 respectively. However, their low precision (0.531 and 0.570 respectively) underperforms the ViLT model. Summarizing images to fine-tune smaller language models tends to result in improved recall albeit at the cost of precision. This approach achieves the highest among the multimodal models with an F1-score of 0.677.

Model	Precision	Recall	F1
Baseline (*support* only)	0.395	1.000	0.566
zero-shot	0.440	0.290	0.350
four-shot	0.420	0.640	0.500
four-shot w/ k-means	**0.450**	**0.700**	**0.550**

Table 3: *Support* stance performance using LLaMA-2 based few-shot experiments.

We compare our few-shot experiments with the baseline *support* only stance predictions to observe that both zero-shot and four-shot models underperform the baseline. The best performance is demonstrated using the four-shot model with k-means clustering. Clustering was found to improve the recall by 6% while precision has improved by 3%. F1-score has improved by 5% to 0.550. Few-shot LLaMA-2 underperforms the ensemble model at stance prediction.

Model	Precision	Recall	F1
ViLT	0.701	0.867	0.775
FLAVA	0.750	0.690	0.720
Multimodal RoBERTa	0.913	0.464	0.615

Table 5: *Oppose* stance performance using image-text multimodal transformer models.

5.2 Oppose Stance

Table 4 shows that the language models have higher precision than recall for the oppose class as compared to the support class (Table 1). Higher precision and lower recall shows us that the text-based language models prioritize predicting the support stance (minority class). Moreover, the ensemble approach outperforms other language models even on the oppose stance. For the multimodal models, both the ViLT and FLAVA models demonstrated superior performance for the oppose class (shown in Table 5) compared to the support class (shown in Table 2). However, the multimodal RoBERTa model follows similar pattern as text-based language models, in terms of scoring high on recall for support class vs oppose class. For LLaMa-2 experiments, The F1 scores for the support class (Table 3) across all methods are consistently higher compared to the oppose class (Table 6). This suggests that LLaMa-2 is more adept at discerning patterns associated with the support class than those of the oppose class.

Model	Precision	Recall	F1
XLNet	0.927	0.630	0.750
Bloom-1B	0.790	0.870	0.830
Bloom-560M	0.919	0.757	0.770
Transformer-XL	0.880	0.569	0.691
DeBERTa-v2	0.770	0.640	0.700
XLM-RoBERTa	0.691	0.899	0.781
Ensemble	**0.929**	**0.796**	**0.857**

Table 4: *Oppose* stance performance using text-based transformer models.

Model	Precision	Recall	F1
Baseline (*Oppose* only)	0.605	1.000	0.754
zero-shot	0.690	0.060	0.110
four-shot	0.770	0.300	0.430
four-shot w/ k-means	0.740	0.270	0.400

Table 6: *Oppose* stance performance using LLaMA-2 based few-shot experiments.

6 Discussion

Popular pre-trained language models such as XLNet, BLOOM, Transformer-XL, DeBERTa-v2 and XLM-RoBERTa were fine-tuned for stance prediction on tweets about gun control and abortion. Results demonstrate that the ensemble of these models performs better than any of the constituent models. However, the disparity is limited. XLNet achieves better recall than the ensemble model and similarly, the BLOOM-560M underperforms the ensemble by 0.026 (though precision of the ensemble is higher by 0.036). This raises the trade-off question between ensemble performance vs. the large computational requirement justified for marginal improvement in the performance.

The best performing multimodal model used the image content summarized as text, unlike its counterpart models that operate in pixel space. We believe the diversity of the images contributes to this difference. In addition to typical images containing people and objects such as guns, trucks and so on, the training set also contained propaganda-related material such as posters with statements. While vision-language models are increasingly getting better at object-centric tasks, understanding such material is closely related to problems such as optical character recognition, which are not often explored in pretraining vision-language models. Our instruction-based image summarization suggests that when explicitly prompted, vision-language models excel not just at object-centric descriptions of images but also at recognizing text from images. Attempts were made to incorporate demographic factors such as number of people in the image, their skin color and gender. However, manual inspection revealed that the resultant instructBLIP predictions were not reliable. Despite augmenting language modality with images in different ways, text-based models outperformed the multimodal models.

Out-of-the-box LLaMA-2 underperforms the baseline *support* only prediction model. However, prompting through four-shot examples greatly improves the performance. This is further enhanced by using in-context examples. This demonstrates that in-context examples that potentially share similar theme (not necessarily the stance) tend to capture the stance better than arbitrary examples from the dataset. The themes were found to include discussions along mental health, effects on children, racism, illegal acquisition, etc. for the gun control dataset; Supreme Court, birth control, religion, reproductive rights, etc. for the abortion dataset.

7 Conclusions and Future Work

Our investigation questions the necessity of images to predict stance in multimodal tweets through different ways of using image-based information in conjunction with text-based language models and investigating the inherent capabilities of LLMs for stance prediction. Results suggest that the best performance can be achieved using an ensemble of language models. Our experiments with multimodal models do not completely refute the utility of images for stance prediction, rather they merely evaluate the current state-of-the-art multimodal models. Incorporating domain knowledge (Lewis et al., 2021), and alternative prompting methods like Question Decomposition (Radhakrishnan et al., 2023) and Tree-of-Thought (Yao et al., 2023) which provide the rationale for the prediction in addition to the stance provide a future direction to address the limited performance with LLaMA-2.

References

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov.

2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. *arXiv preprint arXiv:2006.00052*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. Imagearg: A multi-modal tweet dataset for image persuasiveness mining. *arXiv preprint arXiv:2209.06416*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Question decomposition improves the faithfulness of model-generated reasoning.

Maud Reveilhac and Gerold Schneider. 2023. Replicable semi-supervised approaches to state-of-the-art stance detection of tweets. *Information Processing & Management*, 60(2):103199.

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*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Parinaz Sobhani. 2017. *Stance detection and analysis in social media*. Ph.D. thesis, Universite d'Ottawa/University of Ottawa.

Thomas Wolf Philipp Schmid Zachary Mueller Sourab Mangrulkar Marc Sun Benjamin Bossan Sylvain Gugger, Lysandre Debut. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`.

Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18:267–276.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A multi-task model for emotion and offensive aided stance detection of climate change tweets. *Proceedings of the ACM Web Conference 2023*.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiaoyong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20:447 – 482.

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. Huggingface's transformers: State-of-the-art natural language processing.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer. 2018. Proceedings of grand challenge and workshop on human multimodal language (challenge-hml). In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*.

Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multi-modal meta multi-task learning for social media rumor detection. *IEEE Transactions on Multimedia*, 24:1449–1459.

A Appendix

This appendix provides details such as the number of parameters in the final classification, hyperparameters and finetuning approach for various models[3] in this work. All models used the Adam (Loshchilov and Hutter, 2017) optimizer.

Model	Size of classification head
XLNet	1024
Bloom-1B	64
Bloom-560M	64
Transformer-XL	1024
DeBERTa-v2	1536
XLM-RoBERTa	768
Multimodal RoBERTa (MLP)	1536
FLAVA	768
ViLT	768

Table 7: Table showing the size of the final classification layer for various models used in this work.

A.1 Ensemble Stance Prediction Model

We employed various pretrained language models, specifically XLNet[4], BLOOM-560M[5], Transformer-XL[6], DeBERTa-v2[7], and XLM-RoBeRTa[8]. Each model was augmented with a classification head for binary sequence classification tasks. The summary of the size of the classification head for each model is provided in Table 7. We utilized Adam optimizer with a learning rate of 1e-3. A learning rate scheduler was also incorporated into the training regimen with a patience of 3. To mitigate the risk of model overfitting, a weight decay parameter was set at 0.01. All models were trained for 10 epochs.

A.2 Multimodal Stance Prediction Model

For the multimodal RoBERTa [9], the learning rate was configured at 5e-2, and the weight decay parameter was set at 0.01 during the fine-tuning process. The training continued until the validation loss ceased to decrease for five consecutive epochs. Figure 1 presented the visualization of the Multimodal RoBERTa. For the ViLT[10] model, a low learning of 2.25e-6 was found to be optimal. The model underwent training for a total of 10 epochs. In the case of the FLAVA[11] model, an early stopping mechanism was implemented, resulting model was trained for six epochs prior to any increase in validation loss. The learning rate for this model was set at 5e-5.

A.3 Few-shot Stance Prediction Model

In this study, we employed the Hugging Face's LLaMa-2 13B[12] model for inference, leveraging the capabilities of Hugging Face Accelerate (Sylvain Gugger, 2022). The experimental design utilized Langchain[13] to formulate a tripartite template for prompt engineering. The template is segmented into three distinct components: The system prompt, which serves as a generic instructional scaffold for the language model, a set of few-shot examples to guide the model's responses, and the test set tweet that the model is tasked to analyze. While the standard convention of using no examples for zero-shot and sampling four arbitrary examples for four-shot prediction was used, in the four-shot with k-means, the training set is initially partitioned into clusters using the k-means algorithm (12 clusters for gun control and 13 for abortion). For each test example, its corresponding cluster is predicted, and four examples are randomly sampled from the clus-

[3] code used in this work would be made available after the review process to preserve the anonymity of the authors

[4] https://huggingface.co/xlnet-base-cased

[5] https://huggingface.co/bigscience/bloom-560m

[6] https://huggingface.co/transfo-xl-wt103

[7] https://huggingface.co/microsoft/deberta-v2-xlarge

[8] https://huggingface.co/facebook/xlm-roberta-xl

[9] https://huggingface.co/roberta-base

[10] https://huggingface.co/docs/transformers/model$_doc/vilt$

[11] https://huggingface.co/facebook/flava-full

[12] https://huggingface.co/meta-llama/Llama-2-13b-hf

[13] https://github.com/langchain-ai/langchain

Gun control	Abortion
Gun violence as a mental health problem	Natural Law Right to Life
Effects of gun violence on children	Abortion is evil
Pro-gun control politicians	Supreme Court and abortion
Racism and gun control	Abortion is murder
Trump and guns	Birth control pills
Illegal acquisition of guns	Pro-life
Supreme Court and gun control	Religion and motherhood
Second amendment right	Reproductive rights of women
	#savethebabyhumans hashtag
	Roe v. Wade abortion case

Table 8: Themes identified using k-means clustering for few-shot examples in gun control and abortion datasets.Same theme(s) captured by multiple clusters resulted in fewer themes than reported clusters.

ter as few-shot examples. The optimal number of clusters was ascertained using the Elbow Method (Thorndike, 1953). Table 8 presents some prominent themes found using k-means clustering in gun control and abortion datasets. For LLM output generation, the temperature parameter was set to zero, and the 'top_k' parameter was configured at 30. We employed a Multinomial sampling strategy, setting the do_sample = True and num_beams parameter to 1. An exemplar of the prompt template employed is depicted in Figure 2.

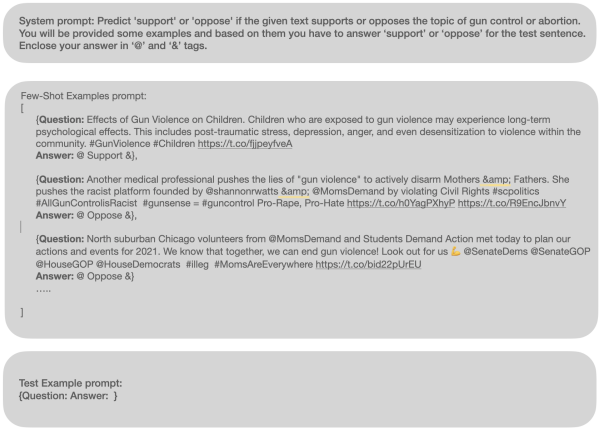

Figure 2: The provided illustration depicts a k-means few-shot prompt template employed in our experimental investigations conducted on the gun control dataset. A comparable configuration was also applied when examining the abortion dataset. For conciseness, we have omitted the inclusion of all four examples in this presentation.