Cross-Domain Persuasion Detection with Argumentative Features

Bagyasree Sudharsan and **Maria Leonor Pacheco**
University of Colorado Boulder
{bagyasree.sudharsan, maria.pacheco}@colorado.edu

Abstract

The main challenge in cross-domain persuasion detection lies in the vast differences in vocabulary observed across different outlets and contexts. Superficially, an argument made on social media will not look like an opinion presented in the Supreme Court, but some of the latent factors that make an argument persuasive are common across all settings. Regardless of domain, persuasive arguments tend to use sound reasoning and present solid evidence, build on the credibility and authority of the source, or appeal to the emotions and beliefs of the audience. In this paper, we show that simply encoding the different argumentative components and their semantic types can significantly improve a language model's ability to detect persuasion across vastly different domains.

1 Introduction

Persuasion is the process of guiding someone to adopt a particular way of thinking or behaving. It is a natural part of everyday life, helping people resolve small disagreements and find common ground. At the same time, it is also a powerful tool used by leaders and institutions to shape society's understanding of broader and more complex issues. Different arguments have different levels of persuasiveness, often determined by specific syntactic and semantic markers (Habernal and Gurevych, 2016a; Ta et al., 2022). Interestingly, this holds true regardless of the context in which arguments are made. In Fig. 1 we present two examples from two different domains (Reddit and Supreme Court proceedings) where the same argumentative strategy was used with the intent of persuading. In both cases, the writer introduced an interpretative claim and offered a premise that appealed to the emotions and beliefs of the audience (known as pathos). In this paper, we build on the idea that these strategies can be applied across domains and use them to detect persuasion in unseen language contexts.

[Welcoming immigrants and refugees has been our country's unfair advantage]$_{claim\_interpretation}$... [As many of you know, I am the son of an undocumented immigrant from Germany and the great grandson of refugees who fled the Armenian Genocide]$_{premise\_pathos}$

[Under the Due Process and Equal Protection Clauses of the 14th Ammendment couples of the same-sex may not be deprived of that right and that liberty]$_{claim\_interpretation}$... [Their hope is not to be condemned to live in loneliness, excluded from one of civilization's oldest institutions. They ask for equal dignity in the eyes of the law]$_{premise\_pathos}$

Figure 1: Similar argumentation strategies in two domains. Reddit post by Alexis Ohanian, co-founder of Reddit (Top). Supreme Court opinion by Justice Anthony Kennedy in Obergefell vs. Hodges (Bottom).

The task of detecting persuasion is not new; numerous studies have examined this problem, primarily with the aim of identifying persuasive strategies to counter misinformation and propaganda (Da San Martino et al., 2020; Nikolaidis et al., 2024). Much of this research has aimed to identify specific techniques in persuasive material with varying degrees of success. Previous studies have examined diverse media, including news articles (Piskorski et al., 2023), social media posts (Tan et al., 2016), legal proceedings (Danescu-Niculescu-Mizil et al., 2012), images (Liu et al., 2023), and memes (Dimitrov et al., 2024). They have also addressed a wide range of domains, from political discourse (Lazer et al., 2018) to medical information (Kamali et al., 2024). However, the vast majority of these studies remain confined to a single domain.

In this paper, we study persuasion detection in a cross-domain setting. Prior work has looked at some aspects of cross-domain transfer in persuasion-related tasks, such as topic-agnostic persuasive dialogue generation (Jin et al., 2024) and cross-lingual variation in persuasive language (Li

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 372–380
November 8-9, 2025

et al., 2024). In the former work, although topics varied, all the conversations followed the same general structure and style; a turn-taking discussion about daily-life situations. In the latter work, although the language varied, all instances were taken from the same media platform. In contrast, our work focuses on exploring whether explicitly modeling argumentation components and their modes of persuasion can improve cross-domain transfer when domains differ along more than one dimension. We build on the premise that argumentation strategies are somewhat domain-agnostic and look at three domains with different purposes, audiences, structures, and argumentation styles. To this end, we make the following contributions.

1. We propose a simple framework for introducing information about the structure of the argument and the persuasion mode for the persuasion detection task.
2. We design a challenging cross-domain experiment using three domains that differ significantly in language use, argument length, and argumentation style: social media, legal proceedings, and formal debates on general topics.
3. We show that regardless of this variation, argument information can significantly improve the ability of fine-tuned language models to detect persuasion across domains.

2 Related Work

Persuasion Detection. There are two main lines of work in the space of persuasion detection: one that frames the problem as a classification task in which the goal is to identify if a text instance is (more or less) persuasive (Habernal and Gurevych, 2016a,b; Dutta et al., 2020; Darnoto et al., 2023), and one that is concerned with identifying the specific persuasion techniques employed in the text (Braca and Dondio, 2023; Dimitrov et al., 2021, 2024; Iyer and Sycara, 2019; Nayak and Kosseim, 2024). Most of these studies employ fully supervised approaches with training data from the target domain. A notable exception is the framework proposed by Iyer and Sycara (2019), which forgoes supervision by relying on syntactic parse trees to identify persuasion tactics. In contrast, we rely on simple signals from the argumentation structure to achieve cross-domain transfer.

Argumentation Mining. There is a large body of work dealing with argumentation mining in the context of persuasive texts. Some studies focus on extracting trees from long documents to represent the overall structure of the arguments made (Stab and Gurevych, 2017; Widmoser et al., 2021). Earlier work has also used argument components to improve high-level persuasion detection (Dutta et al., 2020), although in single-domain scenarios. Chakrabarty et al. (2019) combines these two areas of work by first extracting structured arguments using rhetorical structure theory and then infusing this information into a language model to identify persuasive online discussions. We follow a similar idea but considerably simplify the way in which we model the argumentation structure.

3 Methodology

In this section, we present the argumentation taxonomy and datasets used, as well as our approach to predict argumentation components and persuasion.

3.1 Argumentation Taxonomy

We build on the taxonomy used by Hidey et al. (2017) to qualify opinions on Reddit. They use two types of argumentation components: *claims* and *premises*. Claims are the main statements or conclusions that are being proven. They represent the key ideas that the writer wants the audience to accept as true. Premises, on the other hand, are the supporting statements that provide the reasons or evidence for the claims. The premises serve as the foundation on which the arguments are built.

Premises are further classified into three semantic types according to their mode of persuasion. These modes include *ethos* (appeals to the writer's character), *logos* (appeals to reason), and *pathos* (appeals to emotions and beliefs). We also consider different combinations of these semantic types. For claims, the types are based on a simplification of the Freeman proposition (Freeman, 2011) - *interpretation*, *rational evaluation*, *emotional evaluation*, *agreement* and *disagreement*.

3.2 Datasets

We choose three vastly different datasets to evaluate cross-domain transfer.

The **Reddit Dataset (CMV)** consists of about 290,000 debate threads from the subreddit r/ChangeMyView (CMV) (Tan et al., 2016). Each argument is marked as successful, unsuccessful, or neutral, based on whether or not the reply in the thread received a delta from the original poster. The "unsuccessful" label is used when an attempt to convince was made but failed. The "neutral" label refers to texts that are not argumentative, that is, they are not trying to convince the listener of anything. These may be unrelated to the topic at hand or may be simple sentences or phrases that provide no additional information, such as "Yes, I agree", or "Have you heard of XYZ?" where XYZ has no bearing on the argument. These categories follow Tan et al. (2016). In their work, the authors exclude neutral comments and analyze only argumentative ones; in contrast, we retain neutral comments to train our sequence labeling classifiers and enhance the transferability of our framework, since neutral content is common in realistic scenarios.

The **Supreme Court Oral Arguments Dataset (SCOA)** consists of about 70,000 arguments from various proceedings in the Supreme Court (SCOA) (Danescu-Niculescu-Mizil et al., 2012). Each argument is marked as successful, unsuccessful, or neutral, based on whether or not it was made by the side that won the case. Neutral arguments here cover, in addition, cases where the outcome is not clear.

The **Anthropic Persuasion Dataset (AP)** contains a little less than 4000 claims on a range of general topics, with arguments generated by both humans and LLMs supporting these claims. Each claim has an initial human rating for how much they agree with the initial statement and a human rating for how much they agree after hearing the argument. An argument is considered successful if a person changes their rating from a "disagree" rating (0-4) to an "agree" rating (>4), or if they change their rating by two or more points. Otherwise, it is considered unsuccessful. As this dataset contains only arguments and no neutral examples, it is used solely to evaluate the transferability of the other two models.

These three datasets differ significantly, making cross-domain transfer challenging. The average length of an argument in SCOA is several orders of magnitude larger than that of an argument in CMV, and the lengths of arguments in AP are more concentrated in between the two. These trends can be observed in Fig. 2. Similarly, we observe relatively low token overlap between the CMV and SCOA datasets (Fig. 3). The style and structure of the arguments in the SCOA dataset are also vastly different, as seen in the example in Fig. 1, with more formal speech and more nouns of address.

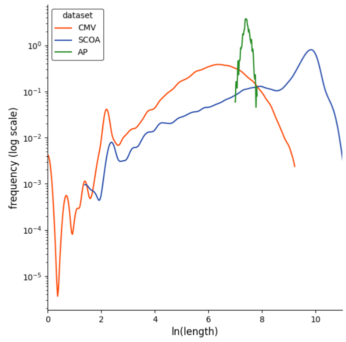

Figure 2: Length Distribution in Datasets

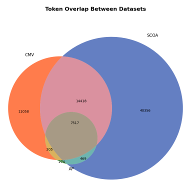

Figure 3: Token Overlap in Datasets

3.3 Model

We propose a simple pipeline that first identifies all argumentation components and their types and then uses this information to predict persuasion.

Identifying Argumentation Components and Their Types. The first step in our approach is to segment the input text and identify the correct argumentation component and semantic type for each segment.

In our implementation, each segment corresponds to a single sentence. Certain conjunctive words tend to indicate the start of a new logical phrase, and so further splitting is done whenever the following are observed: 'but', 'because', 'therefore', 'thus', 'hence', 'however', and 'since'. We break the task into three classification sub-tasks: classifying segments into argumentation components (claims, premises, none), predicting the semantic type of claim (interpretation, rational evaluation, emotional evaluation, agreement, disagreement), and predicting the semantic type of premise (ethos, logos, pathos). All sub-tasks are formulated as sequence-labeling tasks using transformer-based language models.

Predicting Persuasion. Once argumentation components have been identified, we turn our attention to predicting whether an argument is persuasive. Following prior work (Tan et al., 2016; Danescu-Niculescu-Mizil et al., 2012), we define this as a multiclass classification task where the label can be one of: *persuasive*, *not persuasive*,

Model	F1
ArgCompClassifier	0.855
ClaimClassifier	0.696
PremiseClassifier	0.650

Table 1: Text Segment Labeling Classifiers on CMV

Model	SCOA F1	AP F1
ArgCompClassifier	0.702	0.564
SemTypeClassifier	0.368	0.230

Table 2: Text Segment Labeling Classifiers on SCOA and AP (Based on Manual Annotations)

neutral. To use the argumentation taxonomy information, we introduce two special tokens at the beginning of each text segment to identify both the component and its type. We use transformer-based language models for this task.

4 Experimental Results

We describe our experiments using the approach outlined in Sec. 3 to predict persuasion across domains, that is, when training on one dataset and predicting on a different dataset. We also evaluate the performance of our argumentation component and semantic type classifiers. In the CMV data set, this evaluation is based on the corpus of annotated argumentation tags, while in the SCOA and AP datasets, it is based on a much smaller subset of 60 samples that were manually annotated solely for this evaluation.

Experimental Settings. For the argumentation component and semantic type classifiers, we fine-tune DistilBERT (Sanh et al., 2020), while for our cross-domain persuasion classifiers, we fine-tune both DistilBERT (results in Appendix C) and RoBERTa (Liu et al., 2019). Each input is truncated to the maximum length allowed by BERT (512). In all cases, we use 4-fold cross-validation. We test the following combinations:

1. *Segment Labels:* Models were trained using only the argumentation component tags for each segment (ArgComps), using both the argumentation component tags and the semantic type tags (SemTypes), and using neither (Baseline).
2. *Training Data* - Each model is trained on CMV data or SCOA data. Randomly selected subsets of 7500 arguments from the CMV and SCOA datasets are used, with an even split between the number of successful, unsuccessful, and neutral arguments.

The evaluation results for the persuasion classifier model are shown in Table 3. We observe a marked improvement in the model's ability to detect persuasiveness when argument information is explicitly encoded, for both the RoBERTa and DistilBERT classifiers (see Appendix C), showing that this method works irrespective of the base model. However, the performance improvement is minimal on the in-domain task, which could be attributed to segment tag identification errors propagated through the argument component and type classifiers (See Tab. 2). Since argumentation component and semantic type tag annotations are only available for a small subset of the CMV dataset (Hidey et al., 2017), the predictions made by these are treated as ground truth tags while training the final models. While these are able to predict argumentation components and their types quite well for unseen text segments from the CMV set, they were not explicitly trained to classify SCOA or AP segments. Therefore, it is likely that there are more errors in identifying the argumentation components for SCOA, which would also explain why the classifiers trained in this data set offer a lower improvement when using the argumentation information.

Nevertheless, the improvements offered on the cross-domain tasks are encouraging and indicate that this approach has potential. These results show that when a pre-trained transformer encounters types of arguments it has not seen before, such as in a different dataset with different style and vocabulary, the sequence tags do indeed help it identify similarities between training data and the testing data and thus make better predictions. We note that for different datasets, different levels of argumentation information prove useful, with the argumentation component sometimes being sufficient to see a gain, and with the semantic types being required in other cases. Additionally, we observe that the performance of the classifiers with argumentative features on the AP dataset is markedly higher despite the absence of "neutral" text segments, which are present in the training data. This further suggests that our method does indeed generalize.

We have chosen to use plain DistilBERT/RoBERTa, without argumentative features, as our baseline. Our goal is to explore whether incorporating explicit argumentative features in a model improves its performance, which we have shown that it does, rather than obtaining state-of-the-art results for persuasion detection. To accurately gauge whether our hypothesis holds

Model	CMV	SCOA	AP
BaselineLlama	0.375	0.134	0.386
LlamaArgComps	0.389	0.100	0.335
BaselineCMV	**0.570** $\pm$ **0.002**	0.311 $\pm$ 0.004	0.292 $\pm$ 0.003
CMVArgComps	0.570 $\pm$ 0.001	**0.377** $\pm$ **0.004**	0.319 $\pm$ 0.000
CMVSemTypes	0.561 $\pm$ 0.000	0.331 $\pm$ 0.003	**0.344** $\pm$ **0.010**
BaselineSCOA	0.172 $\pm$ 0.001	0.689 $\pm$ 0.001	0.263 $\pm$ 0.001
SCOAArgComps	**0.197** $\pm$ **0.002**	**0.698** $\pm$ **0.000**	**0.310** $\pm$ **0.002**
SCOASemTypes	0.194 $\pm$ 0.002	0.686 $\pm$ 0.000	0.272 $\pm$ 0.003

Table 3: Persuasion Detection Results Across Domains - RoBERTa and Llama3.1

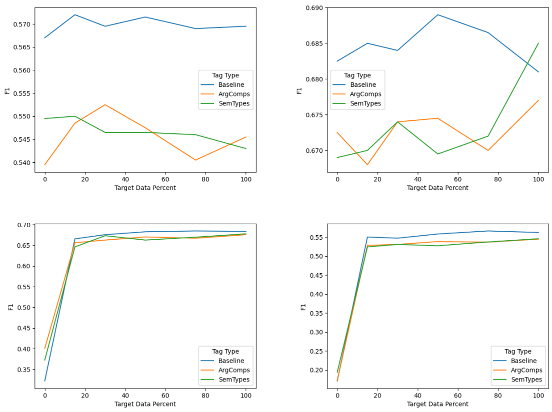

Figure 4: Performance (F1 Score) as more target data is added during training. From left to right: (a) Train on CMV, Eval on CMV (b) Train on SCOA, Eval on SCOA (c) Train on CMV, Eval on SCOA (d) Train on SCOA, Eval on CMV

for other persuasion detection methods, we would potentially have to explore different injection strategies. For completeness, we also include a simple LLM baseline.

LLM Baselines. The transformer-based results are comparable to two-shot LLM baselines: simple prompts passed to a Llama3.1 model, with and without argumentation information. These are visible in Tab. 3.

Training with Target Data. We also performed further experiments to analyze how the performance of these models changed as they saw more training data from the target domain, presented in Figure 4. It is apparent that persuasion detection is a challenging task; all models eventually plateau in performance and are unable to cross a certain threshold.

5 Conclusion and Future Work

While the proposed approach is simple, our findings represent a solid proof of concept that adding inductive bias in the form of argumentative structures and modes of persuasion significantly improves cross-domain persuasion detection, even when this information is noisy. Further, these findings hold for domains with substantial differences in purpose, vocabulary, text length, and style.

In future work, using a more comprehensive scheme to represent arguments, such as Walton's argumentation scheme, could potentially provide a richer representation of the latent structure of the argument. Additionally, other ways of incorporating this information could be explored, such as by cross-attending transformer-based representations of text and graphical networks that model relations between argument components explicitly (Hua et al., 2023), or by combining language model inferences with probabilistic logical inference (Quan et al., 2024; Nafar et al., 2024).

While certain argumentative strategies have been proven to be more effective in certain situations or with certain people (Wang et al., 2019), we have shown the impact of including some underlying indicators that are useful to gauge the persuasiveness of an argument in different domains. These can be further improved by including user- or context-specific information.

Limitations

Our work has two main limitations. Firstly, the scope of our study is small. While we use three datasets from three different domains, these datasets do not cover the full range of domains where persuasion is of importance. A larger study could further verify that our findings hold for other classes of domain variations such as topic and language. Secondly, the proposed approach is somewhat dependent on how well we can identify argument components and their types. We showed that even a noisy representation is beneficial, as we did not fully verify the validity of the intermediate representations for the SCOA case, and still saw an improvement. However, further evaluation is needed to quantify the noise-to-performance ratio.

Regardless of these limitations, we believe that our findings constitute a meaningful, focused contribution that could inform future work in cross-domain persuasion detection.

References

Annye Braca and Pierpaolo Dondio. 2023. Persuasive communication systems: a machine learning approach to predict the effect of linguistic styles and persuasion techniques persuasive communication systems. *Journal of Systems and Information Technology*, 25:1328–7265.

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, Hong Kong, China. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, page 699–708.

Brian Rizqi Paradisiaca Darnoto, Daniel Siahaan, and Diana Purwitasari. 2023. Automated detection of persuasive content in electronic news. *Informatics*, 10(4).

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Sebastian Duerr and Peter A. Gloor. 2021. Persuasive natural language generation - a literature review. *ArXiv*, abs/2101.05786.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085.

James B Freeman. 2011. *Argument Structure:*, 2011 edition. Argumentation Library. Springer, Dordrecht, Netherlands.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Rahul Radhakrishnan Iyer and Katia P. Sycara. 2019. An unsupervised domain-independent framework for automated detection of persuasion tactics in text. *ArXiv*, abs/1912.06745.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Danial Kamali, Joseph D. Romain, Huiyi Liu, Wei Peng, Jingbo Meng, and Parisa Kordjamshidi. 2024. Using persuasive writing strategies to explain and detect health misinformation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17285–17309, Torino, Italia. ELRA and ICCL.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136, Denver, CO. Association for Computational Linguistics.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024. Uncovering differences in persuasive language in Russian versus English Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 21–35, Miami, Florida, USA. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

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*, pages 120–132, Singapore. Association for Computational Linguistics.

Aliakbar Nafar, K. Brent Venable, and Parisa Kordjamshidi. 2024. Teaching probabilistic logical reasoning to transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1615–1632, St. Julian's, Malta. Association for Computational Linguistics.

Kota Shamanth Ramanath Nayak and Leila Kosseim. 2024. Analyzing persuasive strategies in meme texts: A fusion of language models with paraphrase enrichment. *Preprint*, arXiv:2407.01784.

CS Ngai and RG Singh. 2020. Relationship between persuasive metadiscoursal devices in research article abstracts and their attention on social media.

Nikolaos Nikolaidis, Jakub Piskorski, and Nicolas Stefanovitch. 2024. Exploring the usability of persuasion techniques for downstream misinformation-related classification tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6992–7006, Torino, Italia. ELRA and ICCL.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise Dennis, and Andre Freitas. 2024. Enhancing ethical explanations of large language models through iterative symbolic refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–22, St. Julian's, Malta. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. Modeling relational data with graph convolutional networks. *Preprint*, arXiv:1703.06103.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

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

Vivian P. Ta, Ryan L. Boyd, Sarah Seraj, Anne Keller, Caroline Griffith, Alexia Loggarakis, and Lael Medema. 2022. An inclusive, real-world investigation of persuasion in language and verbal behavior.

Vivian Ta-Johnson, Ryan Boyd, Sarah Seraj, Anne Keller, Caroline Griffith, Alexia Loggarakis, and Lael Medema. 2022. An inclusive, real-world investigation of persuasion in language and verbal behavior. *Journal of Computational Social Science*, 5.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings* *of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Manuel Widmoser, Maria Leonor Pacheco, Jean Honorio, and Dan Goldwasser. 2021. Randomized deep structured prediction for discourse-level processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1174–1184, Online. Association for Computational Linguistics.

A Appendix - LLM Prompts

Llama Prompt Without Argument Information: "I will give you a paragraph of text containing an argument that is trying to be persuasive. Analyze the argument structure and the argumentative strategies employed, and using this, classify the argument as SUCCESSFUL, UNSUCCESSFUL or NEUTRAL. If it a strong argument that is likely to succeed at convincing someone, it is SUCCESSFUL, otherwise it is UNSUCCESSFUL. If it not an argument, say NEUTRAL. Return one of: SUCCESSFUL, UNSUCCESSFUL, NEUTRAL. Some examples to illustrate:
Input: <input1>
Expected Output: <output1>
Input: <input2>
Expected Output: <output2>
Give similar outputs for the argument below:
<text>"

Llama Prompt With Argument Information: "I will give you a paragraph of text containing an argument that is trying to be persuasive. Analyze the argument structure and the argumentative strategies employed, and using this, classify the argument as SUCCESSFUL, UNSUCCESSFUL or NEUTRAL. If it is a strong argument that is likely to succeed at convincing someone, it is SUCCESSFUL, otherwise it is UNSUCCESSFUL. If it not an argument, say NEUTRAL. This paragraph will also have special tags, enclosed in '[]', which states whether the following text segment is a claim, premise, or is neutral, followed by a tag stating the type of claim or premise. Use this information to make your decision. Return one of: SUCCESSFUL, UNSUCCESSFUL, NEUTRAL. Some examples to illustrate: Input: <input1>
Expected Output: <output1>
Input: <input2>
Expected Output: <output2>
Give similar outputs for the argument below:
<text>"

The above prompts are populated with each sample in the test set, and the examples are entered in the input and output spaces. The inputs in the prompt with the argument information also contain argument component and semantic type tags.

B Appendix - Hyperparameters

Information for all hyper-parameters used can be observed in Tab. 4.

C Appendix - DistilBERT Cross-Domain Persuasion Classifier

The results for the persuasion classifier trained using DistilBERT can be found in Tab. 5. There is a slight downturn in performance on the in-domain task, but an improvement in the cross-domain task, as discussed above.

Model	Number of Epochs	Training Batch Size	Optimizer	Learning Rate
ArgCompClassifier	3	8	AdamW	3e-5
ClaimClassifier and PremiseClassifier	5	8	AdamW	3e-5
Baseline Persuasion Classifiers	3	16	AdamW	3e-5
Persuasion Classifiers with Argumentation Info	5	8	AdamW	3e-5

Table 4: DistilBERT and RoBERTa Fine-Tuning Hyperparameters

Model	CMV	SCOA	AP
BaselineCMV	**0.567±0.005**	0.322±0.000	0.314± 0.001
CMVArgComps	0.539±.004	**0.400±.004**	0.316±0.05
CMVSemTypes	0.551±.002	0.375±.002	**0.351±0.003**
BaselineSCOA	0.168±.002	**0.682±.003**	0.192±0.06
SCOAArgComps	0.180±.004	0.672±.000	0.215±0.005
SCOASemTypes	**0.205±.001**	0.669±.001	**0.256±0.002**

Table 5: Persuasion Detection Results Across Domains - DistilBERT