Reasoning Under Distress: Mining Claims and Evidence in Mental Health Narratives

Jannis Köckritz[1,2], Bahar İlgen[1], Georges Hattab[1,2]
[1] Center for Artificial Intelligence in Public Health Research (ZKI-PH),
Robert Koch Institute, Berlin, 13353, Germany
[2] Department of Mathematics and Computer Science,
Freie Universität Berlin, Berlin, 14195, Germany
Correspondence: KoeckritzJ@rki.de

Abstract

This paper explores the application of argument mining to mental health narratives using zero-shot transfer learning. We fine-tune a BERT-based sentence classifier on ∼15k essays from the Persuade dataset—achieving 69.1% macro-F1 on its test set—and apply it without domain adaptation to the CAMS dataset, which consists of anonymized mental health–related Reddit posts. On a manually annotated gold-standard set of 150 CAMS sentences, our model attains 54.7% accuracy and 48.9% macro-F1, with evidence detection (F1 = 63.4%) transferring more effectively than claim identification (F1 = 32.0%). Analysis across expert-annotated causal factors of distress shows that personal narratives heavily favor experiential evidence (65–77% of sentences) compared to academic writing. The prevalence of evidence sentences, many of which appear to be grounded in lived experiences, such as descriptions of emotional states or personal events, suggests that personal narratives favor descriptive recollection over formal, argumentative reasoning. These findings underscore the unique challenges of argument mining in affective contexts and offer recommendations for enhancing argument mining tools within clinical and digital mental health support systems.

1 Introduction

Argument mining (AM) has produced strong results with structured texts, such as persuasive essays and legal documents (Stab and Gurevych, 2017; Habernal et al., 2023; Lippi and Torroni, 2016). Recent approaches have expanded the scope of AM to include less formal domains, such as online forums and social media (Schaefer and Stede, 2020). However, mental health narratives—personal accounts of psychological distress shared on peer support platforms—remain understudied despite their potential to reveal how individuals think about their mental state (Iskender et al., 2021).

This paper investigates whether models trained on structured, formal-domain texts can be used to analyze mental health narratives in a zero-shot transfer setting. Additionally, we examine how argumentative structures vary across expert-annotated causal factors of distress in mental health–related posts. Our approach involves fine-tuning BERT on the Persuade corpus of argumentative essays (Crossley et al., 2022), applying it without adaptation to the CAMS dataset of Reddit posts about mental health (Garg et al., 2022), and evaluating both transfer performance and shifts in argumentative patterns[1].

Our contributions are:

1. **An empirical evaluation of zero-shot AM across domains** was conducted, demonstrating a decrease in macro-F1 scores from 69.1% in the source domain to 48.9% in mental health narratives and quantifying transfer limitations.

2. **The analysis of argumentative structures in mental health discourse** revealed that personal narratives predominantly consist of experiential evidence (65-77%), with minimal explicit claims. This contrasts sharply with academic writing.

This study deepens our understanding of how argumentation occurs under psychological distress. It also paves the way for the development of domain-specific argumentation management (AM) tools for affective contexts, such as clinical and digital mental health applications.

[1]All code and datasets used in this study are publicly available at https://github.com/JannisKoeckritz/ReasoningUnderDistress

Proceedings of the 12th Argument Mining Workshop, pages 189–194

2 Related Work

Argument mining identifies components such as claims and premises within a text (Lawrence and Reed, 2020). Earlier work demonstrated strong performance in formal domains—persuasive essays (Wachsmuth et al., 2016; Stab and Gurevych, 2017), legal reasoning (Habernal et al., 2023), and debates (Lippi and Torroni, 2016)—where arguments align with clear schemas (Lauscher et al., 2018; Cohan et al., 2019). More recent research has extended AM to informal genres, including online discussions and social media, where models confront implicit argumentation and emotionally charged content (Dusmanu et al., 2017; Vecchi et al., 2021; Cabessa et al., 2024; Mezza et al., 2024). Some studies such as Gupta et al. (2024) propose novel zero-shot methods for argument explication using large language models (LLMs). These LLMs decompose informal arguments into structured components, such as claims, reasons, and warrants.

Research in mental health NLP has focused on diagnostics, such as identifying depression and suicide risk, using lexical and affective features (Malgaroli et al., 2023; Montejo-Ráez et al., 2024), with little attention to argumentative structure. A small number of studies have applied AM to subjective or health-related narratives (Mayer et al., 2020), but cross-domain transfer remains largely unexplored.

This work bridges the fields of argumentation mining (AM) and mental health by applying a formal-domain AM model to CAMS. This reveals the challenges of mining arguments in affect-laden, informal texts. Building on this research, we explore whether formal-domain AM models can be applied to mental health discourse in a zero-shot setting.

3 Data & Methodology

3.1 Datasets

We use two datasets. The Persuade dataset is used to train argument mining models, and the CAMS dataset is used for zero-shot evaluation in the mental health domain.

The **Persuade dataset** (Crossley et al., 2022) contains argumentative essays from U.S. students in grades 6-12 with professional annotations across seven categories: Lead, Position, Claim, Counterclaim, Rebuttal, Evidence, and Concluding Statement. We have consolidated these into three categories: (1) *Claim* (combining original Claim, Counterclaim, and Rebuttal), (2) *Evidence*, and (3) *Other* (consolidating Lead, Position, Concluding Statement, and unannotated text). This simplified taxonomy makes it easier to transfer to informal contexts while preserving the core argumentative distinctions. The dataset consists of approximately 25,000 documents. For training our sentence-level classification model, we used only 15,000 of these documents, corresponding to around 300,000 sentences.

The **CAMS dataset** (Garg et al., 2022) comprises 5,051 Reddit posts that have been annotated for an interpretable causal analysis of mental health issues. It includes 3,155 posts that were crawled from the *r/depression* subreddit, as well as 1,896 reannotated posts from the existing SDCNL dataset. Each post is labeled with one of six categories reflecting psychological distress: (i) no reason, (ii) bias or abuse, (iii) jobs and careers, (iv) medication, (v) relationships, and (vi) alienation. The distribution of posts across causal categories is shown in Figure 1. Trained student annotators performed the annotations following expert-developed guidelines, and a clinical psychologist and a rehabilitation counselor later verified them. Posts were selected using keyword filtering and language criteria to ensure relevance and consistency.

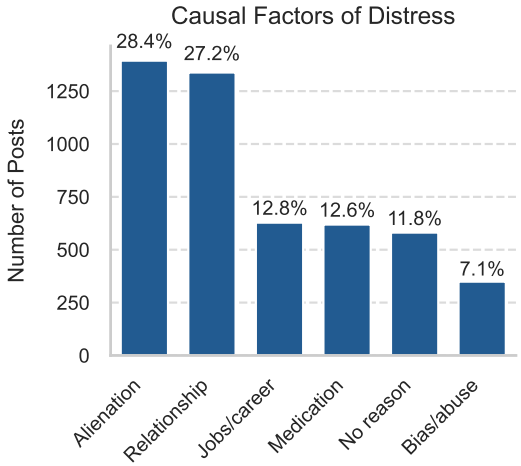

Figure 1: Distribution of Causal Factors in Mental Health Distress This figure shows how 4,963 social media posts are distributed across six categories of causal factors of mental health distress. Alienation (28.4%) and relationship issues (27.2%) collectively account for over half of all posts, highlighting the predominance of social and existential concerns. The remaining categories—jobs/career (12.8%), medication (12.6%), no identified reason (11.8%), and bias/abuse (7.1%)—represent a more diverse set of external stressors and unexplained distress.

3.2 Model Architecture and Training

We fine-tuned a sentence-level classification model based on BERT-base, treating each sentence as a discrete argumentative unit. The preprocessing pipeline involves sentence segmentation using spaCy, followed by BERT tokenization. The model architecture consists of a classification head that receives the token representation from BERT, applies dropout with a probability of 0.1, and passes it through a linear layer with softmax activation to produce the final classification. We use cross-entropy loss with class weighting based on inverse frequency to address class imbalance for training. We also use a batch size of 32, and a learning rate of 5e-4 with linear decay. We employed early stopping based on the validation F1 score with a patience of 20 epochs. The trained model achieves 69.0% accuracy and 69.1% macro F1 on the Persuade test set, performing best on *Evidence* identification (F1: 76.9%).

3.3 Evaluation Methodology

To establish a reliable evaluation baseline for the mental health domain, we manually annotated 150 sentences, randomly sampled from 30 CAMS posts and balanced across six causal-factor categories. Although this subset is a small part of the full CAMS corpus, it was carefully chosen to include a variety of distress sources, making it a good sample for our analysis. Two annotators, both experts in argument mining and mental health discourse, independently labeled each sentence in the CAMS sample as *Claim*, *Evidence*, or *Other*, based on our consolidated taxonomy. The six causal factors were pre-existing annotations in the CAMS dataset; however, the argumentative role labels introduced in this study were newly assigned by the annotators. The annotation process achieved an inter-annotator agreement of Cohen's $\kappa = 0.71$. Disagreements resolved through discussion to create the final gold standard.

We evaluated zero-shot transfer by applying the Persuade-trained model directly to the CAMS dataset without additional training. For the quantitative evaluation, we report the accuracy and F1 scores on the gold-standard subset. Then, to analyze domain-specific patterns, we apply the model to the full CAMS dataset and examine the distribution of argumentative elements across causal factors. Finally, we compare these patterns to those observed in the Persuade corpus.

4 Results

We consolidated the original Persuade annotation scheme by mapping Claim, Counterclaim, and Rebuttal into a single *Claim* category, maintaining *Evidence* as a separate category, and grouping Lead, Position, Concluding Statement, and unannotated text as *Other*.

Our BERT-based sentence classification model achieved an overall accuracy of 69.0% and a macro-averaged F1 score of 69.1% on the Persuade test set. Performance varied across argument categories, with *Evidence* sentences achieving the highest F1 score (76.9%), while Claims proved more challenging (F1: 53.1%). Table 2 presents the detailed performance metrics.

Category	Precision	Recall	F1 Score
Other	0.686	0.642	0.663
Claim	0.578	0.491	0.531
Evidence	0.736	0.806	0.769
Accuracy		0.690	
Avg F1		0.691	

Table 1: Classification performance metrics for the BERT sentence classification model on the Persuade test set.

The confusion matrix (Figure 2) reveals that the model most frequently confused *Claims* with *Evidence* (2,185 instances), indicating the challenge of distinguishing between these categories. *Claims* were also frequently misclassified as *Other* (1,029 instances). The model demonstrated strongest performance in identifying *Evidence*, correctly classifying 13,239 instances.

4.1 Zero-Shot Domain Transfer Evaluation

To evaluate cross-domain generalization, we manually annotated a gold-standard subset of 150 sentences from the CAMS dataset using our three-category scheme. When evaluated against this standard, our model achieved an accuracy of 54.7% and a macro F1 score of 48.9%. Performance varied across categories, with *Evidence* again being most reliably identified (F1: 63.4%), followed by *Other* (F1: 51.3%), while *Claim* classification remained challenging (F1: 32.0%). Compared to the source domain, this represents a 14.3-percentage-point drop in accuracy and a 20.2-point drop in macro F1, highlighting the challenges of cross-domain transfer to mental health narratives. These scores are informative but should be interpreted cautiously

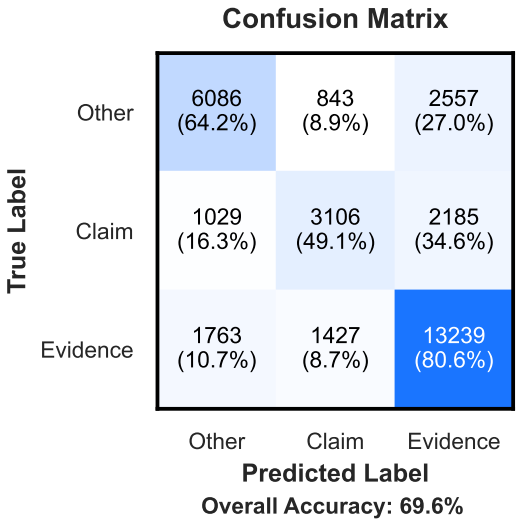

Figure 2: Confusion matrix for the BERT sentence classification model on the Persuade test set, showing the distribution of predicted vs. true labels.

due to the small evaluation sample size and the very low number of claim-labeled sentences. Larger annotated samples are needed to reliably estimate cross-domain generalization, particularly for underrepresented argument types.

Category	Precision	Recall	F1 Score
Other	0.547	0.482	0.513
Claim	0.376	0.279	0.320
Evidence	0.618	0.651	0.634
Accuracy		0.547	
Avg F1		0.489	

Table 2: Zero-shot transfer performance on the gold standard CAMS dataset.

The prevalence of evidence sentences, many of which appear to be grounded in lived experiences, such as descriptions of emotional states or personal events, suggests that personal narratives favor descriptive recollection over formal argumentative reasoning. During the annotation process, sentences were labeled as evidence if they served a justifying function, typically through descriptions of lived experiences, emotional states, or contextual details, even if they lacked external citations. This differs from academic domains, where evidence often consists of formally structured reasoning or references to facts.

A comparative analysis revealed that *Evidence* identification transferred relatively well across domains, while *Claim* recognition showed more significant degradation. This pattern aligns with our hypothesis that personal narratives express claims differently than academic writing does, while the presentation of evidence (often through personal experiences or references to external sources) shows more structural consistency across domains.

4.2 Distribution of Argumentative Elements by Causal Factors of Distress

We used our model to analyze how argumentative elements are distributed across different mental health categories in the CAMS dataset (Figure 3). The analysis reveals distinct patterns across categories. *Evidence* represents the most significant proportion in most categories, accounting for approximately 42-75% of sentences. In contrast, *Claims* remain consistently low across all categories (under 2%). This differs markedly from the Persuade corpus, where *Claims* represent approximately 28% of sentences.

The *Other* class is also well-represented, especially in the "No reason" and "Alienation" categories, where it accounts for about 40-55% of the sentences.

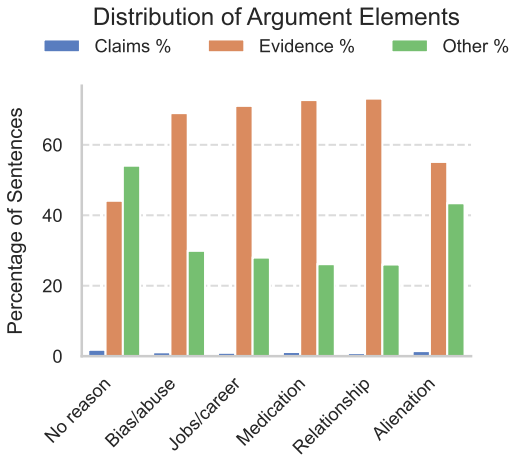

Figure 3: Distribution of argumentative elements (*Claims*, *Evidence*, *Other*) across the expert-annotated causal factors of distress in the CAMS dataset. Evidence statements are prevalent across most categories, while claims are rare.

The prevalence of *Evidence* across all causal factors suggests that personal narratives prioritize experiential descriptions over explicit claims or conclusions. Distress attributed to "Bias/abuse," "Jobs/career," "Medication," and "Relationship" shows the highest proportion of *Evidence* (>60%), indicating a greater focus on descriptive content. The "No reason" and "Alienation" categories exhibit slightly lower *Evidence* proportions and

higher *Other* content, potentially reflecting more abstract or emotional expressions that fall outside our argumentative framework.

5 Discussion

Our results demonstrate the challenges and insights gained from applying argument mining to mental health narratives. The uneven transfer of argumentative components—with *Evidence* transferring more successfully than *Claims*—reveals fundamental differences in how arguments manifest across domains. This asymmetry, coupled with the substantial performance degradation in zero-shot transfer (20.2 point drop in macro F1), highlights the domain-specific nature of argumentative structures.

Mental health narratives exhibit a distinctive argumentative profile: *Evidence* statements (65-77%) dominate across all causal factors of distress. In contrast, *Claims* represent only 1-2% of the content, which is dramatically different from academic writing where claims form the backbone of argumentation. This suggests that, when explaining psychological distress, individuals prioritize experiential descriptions over explicit claim-making, regardless of the attributed cause. The boundary between personal experience (*Evidence*) and interpretation (*Claim*) often blurs in mental health narratives, creating inherent ambiguity. For example, a sentence such as "I stopped going to work because I couldn't get out of bed" can be both a factual recounting and an implied assertion of a causal link. This interpretive ambiguity suggests the need for more nuanced annotation schemes in emotionally charged contexts. These challenges highlight the potential benefits of redefining argumentation categories for mental health discourse.

Although zero-shot classification is simple to implement, it fails to account for domain-specific patterns. More promising approaches include few-shot learning with minimal in-domain data and domain-adversarial training, which explicitly models cross-domain differences. This work contributes valuable insights into cross-domain argument mining for mental health narratives; however, certain limitations should be acknowledged. First, the annotated CAMS subset is relatively small, which may affect generalizability. Additionally, although BERT provides a robust and well-established baseline, future studies could examine more recent transformer models, such as DeBERTa and RoBERTa, as well as instruction-tuned LLMs. Other promising directions include few-shot adaptation, discourse-level modeling, and developing domain-specific taxonomies suited to affective contexts.

Future research should develop argumentation schemes specific to mental health and expand beyond sentence-level classification to capture multi-sentence argumentative structures. Dialog-based systems that integrate interaction and explanation could provide additional value. For instance, Castagna et al. (2023) propose EQRbot, a chatbot that uses expert knowledge to provide argument-based explanations and critical questions. Such systems not only classify argument types but also clarify reasoning—particularly valuable in emotionally charged, ambiguous contexts like mental health discourse. Integrating dialogic and explanatory elements into future AM models could better align computational processing with real-world needs in digital mental health, enhancing clinical applications, peer support, content moderation, and research

6 Conclusion

This study examined the zero-shot transfer of argument mining from structured essays to mental health narratives. Our results show that evidence transfers reasonably well across domains, but claims are more difficult due to how they manifest in emotional contexts. The 20.2-point drop in macro F1 score between domains underscores the need for argument mining techniques tailored to mental health discourse. Promising directions include few-shot learning and domain-adaptive approaches to better capture argumentative structures in narratives about psychological distress. Recent work has emphasized the growing role of AI in public health infrastructure and decision support systems, particularly through explainable and human-in-the-loop approaches to foster trust and transparency (Hattab et al., 2025). Our findings underscore the importance of domain-specific natural language processing (NLP) techniques for understanding patient-generated narratives in digital health contexts.

References

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2024. Argument mining in BioMedicine: Zero-shot, in-context learning and fine-tuning with LLMs. In *Proceedings of the 10th Italian Conference on Com-*

putational Linguistics (CLiC-it 2024), pages 122–131, Pisa, Italy. CEUR Workshop Proceedings.

Federico Castagna, Alexandra Garton, Peter McBurney, Simon Parsons, Isabel Sassoon, and Elizabeth I. Sklar. 2023. Eqrbot: A chatbot delivering eqr argument-based explanations. *Frontiers in Artificial Intelligence*, 6.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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)*. Association for Computational Linguistics.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322.

Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6387–6396, Marseille, France. European Language Resources Association.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276, Bangkok, Thailand. Association for Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

Georges Hattab, Christopher Irrgang, Nils Körber, Denise Kühnert, and Katharina Ladewig. 2025. The way forward to embrace artificial intelligence in public health. *American Journal of Public Health*, 115(2):123–128.

Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. 2021. *Argument Mining in Tweets: Comparing Crowd and Expert Annotations for Automated Claim and Evidence Detection*, page 275–288. Springer International Publishing.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, page 40–46. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1).

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

Stefano Mezza, Wayne Wobcke, and Alan Blair. 2024. Exploiting dialogue acts and context to identify argumentative relations in online debates. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 36–45, Bangkok, Thailand. Association for Computational Linguistics.

Arturo Montejo-Ráez, M. Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

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

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

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, Osaka, Japan. The COLING 2016 Organizing Committee.