Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study

Greta Damo, Elena Cabrio, Serena Villata

Effectiveness of Counter-Speech against Abusive Content: A Multidimensional Annotation and Classification Study

Greta Damo
Université Côte d'Azur
Inria, CNRS, I3S
Sophia Antipolis, France
greta.damo@univ-cotedazur.fr

Elena Cabrio
Université Côte d'Azur
Inria, CNRS, I3S
Sophia Antipolis, France
elena.cabrio@univ-cotedazur.fr

Serena Villata
Université Côte d'Azur
Inria, CNRS, I3S
Sophia Antipolis, France
serena.villata@univ-cotedazur.fr

***Abstract*—Counter-speech (CS) is a key strategy for mitigating online Hate Speech (HS), yet defining the criteria to assess its effectiveness remains an open challenge. We propose a novel computational framework for CS effectiveness classification, grounded in linguistics, communication and argumentation concepts. Our framework defines six core dimensions – Clarity, Evidence, Emotional Appeal, Rebuttal, Audience Adaptation, and Fairness – which we use to annotate 4,214 CS instances from two benchmark datasets, resulting in a novel linguistic resource released to the community. In addition, we propose two classification strategies, multi-task and dependency-based, achieving strong results (0.94 and 0.96 average F1 respectively on both expert- and user-written CS), outperforming standard baselines, and revealing strong interdependence among dimensions.**

Content warning: this paper contains unobfuscated examples some readers may find offensive.

***Index Terms*—Natural language processing, Hate speech, Counter-speech generation, Deep learning, Fairness, Linguistics**

I. Introduction

The rise of social media has intensified concerns over online Hate Speech (HS). Counter-speech (CS) has emerged as an effective strategy for mitigating harmful online discourse. It involves non-aggressive responses, that use credible evidence, factual arguments, and alternative perspectives to challenge hate speech [1], [2]. Counter-speech has been studied and approached through various lenses: some see it as direct responses to hate speech [3], [4], while others argue it should condemn hate, challenge and offer alternative viewpoints, and offer support to the victim [5]–[8]. CS is also valued for its potential to shift beliefs and reduce future hate [9], [10]. CS can be written by experts, online users, or chatbots (based on Large Language Models – LLMs). Expert-written CS typically presents structured reasoning and appeals to shared human values, whereas user- and LLM-generated responses, in particular, tend to be more generic and emotional, and often lacks depth and relies on surface-level rebuttals [11]–[13].

This study evaluates expert- and user-written CS effectiveness using two benchmark datasets [14], [15].This allows us to compare how the origin of a counter-speech message influences its effectiveness. Figure 1 shows an example of a hateful message with user- and expert-written CS responses. Current evaluations of counter-speech effectiveness rely on surface-level metrics – coherence, relevance, or non-offensiveness – assuming well-formed CS is inherently effective [4], [11], [16]–[18]. But concentrating on CS surface form only is not enough to achieve shifting attitudes or reducing harm. Assessing CS effectiveness is crucial, as it has the potential for significant societal impact, by reducing online hostility and potentially curbing offline violence and discrimination.

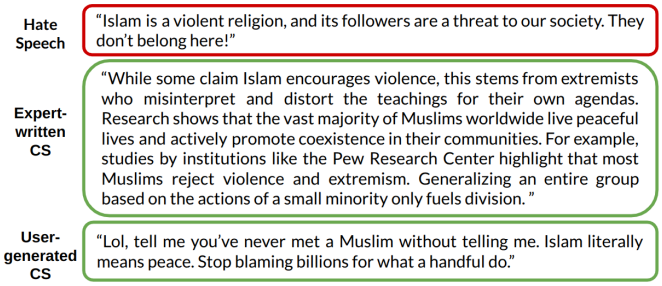

Fig. 1. Examples of expert-written and user-written counter-speech.

To address this gap, this study introduces a novel structured framework grounded in linguistics, communication, rhetoric, and argumentation theories, helping to evaluate counter-speech effectiveness. Our contribution is twofold: **(1) a new multi-dimensional framework for automatic counter-speech effectiveness classification** across six key dimensions – Clarity, Evidence, Emotional Appeal, Rebuttal, Audience Adaptation, and Fairness – from linguistics, communication, rhetoric and argumentation theories, and two classification strategies, **multi-task** and **dependency-based** modeling, that outperform standard baselines, achieving average F1 scores of 0.94 and 0.96 respectively, when combining user-written and expert-written CS data. **(2) a novel human annotated resource** comprising two expert- [15] and user-written [14] counter-speech datasets, enriched with the annotation layer with the six dimensions of our effectiveness framework[1]

[1]Datasets, guidelines, and code are available on GitHub. and HuggingFace.

The remainder of the paper is structured as follows. Section II reviews prior studies on counter-speech and its evaluation. Section III defines the six dimensions used to assess CS effectiveness, tailoring them to existing linguistics and argumentation theories. Section IV details the datasets used for the annotation and the classification task, and the annotation process together with the Inter-Annotator-Agreement results. Section V outlines the models used, and the experimental setup. Section VI presents the results of our classification task, showing comparisons between the baselines and our proposed models. Section VII identifies and discusses challenges in specific dimensions and limitations. Finally, Section VIII summarizes key findings and future directions.

II. Related Work

Research on counter-speech has gained increasing attention as a strategy for combating online hate speech. However, most studies have focused on hate speech detection and counter-speech generation, while the systematic evaluation of CS effectiveness remains under-explored. In this section, we review key contributions on counter-speech analysis, emphasizing the need for a comprehensive CS evaluation framework.

A. Hate Speech & Counter Speech Detection

HS detection is a well-studied NLP task, initially tackled as binary classification on English datasets [19], [20], later extended to fine-grained labels for discrimination and microaggressions [21]–[26]. Early approaches relied on traditional Machine Learning [27], [28], while recent methods leverage transformers like BERT for better contextual modeling [29]–[32].

On the other hand, counter-speech, introduced as a mitigation strategy by [33], has not yet been extensively explored from a detection standpoint. Among the studies, [34] propose an early HS/CS detection pipeline, by applying an ensemble learning algorithm to classify HS and CS tweets, achieving F1 scores between 0.76 and 0.97, emphasizing counter-speech complexity due to subjectivity and definitional ambiguity.

B. Counter Speech Classification & Evaluation

Concerning counter-speech classification, [35] find that a one-shot prompted LLM achieves promising accuracy in classifying manually labeled CS strategies. [36] analyze HS/CS tweet pairs, using a boosting algorithm with TF-IDF and lexical features, which achieves 0.77 F1. [3] also analyze CS and neutral comments on YouTube, finding that TF-IDF vectors combined with logistic regression achieve 0.73 F1. [37] conducts a qualitative analysis of counter-speech examples, while [38] analyze tweets during the COVID-19 crisis, achieving an F1 score of 0.49 for CS.

A key gap in the literature is the lack of standardized evaluation methods for assessing human- or machine-generated counter-speech. Several studies propose evaluation criteria, but no unified framework emerged. [39] assess CS effectiveness through human judgments, focusing on comfort for targets and bystander empathy. [16] and [17] similarly rely on human ratings based on qualities such as informativeness, coherence, and stance. [4] evaluate CS suitability using offensiveness and stance, while [40] propose an LLM-based evaluation framework, grounded in NGO guidelines, which shows strong alignment with human judgments. [41] introduce a conversation-level incivility metric, arguing that CS should be judged by its downstream impact on discourse. Finally, [42] focus on evaluating the human-likeness of AI-generated counter-speech.

While prior work shows growing interest in CS evaluation, we advance beyond previous studies by introducing a human-annotated resource based on our framework that captures fine-grained dimensions of structural coherence, content strength, rhetorical impact, and linguistic strategies, to assess its effectiveness. We also develop classification methods that leverage this framework.

III. Effectiveness Metrics

We propose a framework to evaluate counter-speech effectiveness across six dimensions: Clarity, Evidence, Emotional Appeal, Rebuttal, Audience Adaptation, and Fairness. These are grounded in communication theory.

- **Clarity** ensures structural coherence and logical flow. Effective counter-speech should be clear and logically structured, so that it is easier for the audience to follow the reasoning and understand the main points [43]–[46].
- **Evidence** ensures that effective counter-speech is supported by relevant evidence and examples that support the claims and make it more compelling to the audience. They should be specific to the topics addressed in the hateful message. Studies indicate that presenting multiple pieces of evidence in an argument, such as statistics or expert testimony, enhances persuasiveness and effectiveness [45], [47]–[49].
- **Emotional Appeal** is a rhetorical strategy, used to evoke empathy, or other emotions in the audience, which can help strengthen the counter-speech impact. The importance of emotions and empathy in designing effective counter-speech has been investigated by numerous studies [6], [45], [50]–[53].
- **Rebuttal** shows how using arguments to anticipate and address potential counterarguments, objections, and opposing views improves credibility and strengthens persuasion [45], [50], [54]–[58].
- **Audience Adaptation** involves tailoring the language of the counter-speech to a specific audience, taking into account their level of linguistic ability and knowledge, thus improving understanding and relatability. From a linguistic point of view, a counter-speech is more effective and understandable if the language employed is similar to the one spoken by the audience [59]–[61].
- **Fairness** promotes the use of appropriate and respectful language, respecting freedom of expression, without censoring or dehumanizing the opposing viewpoint. Fair language supports ethical discourse and improves effectiveness [62], [63].

TABLE I
ANNOTATION GUIDELINES FOR CS EFFECTIVENESS.

Categorical Dimensions (1–3 Likert Scale)	
Clarity	3: Clear, logically structured, directly addresses the HS topic. 2: Mostly clear or slightly generic. 1: Vague or incoherent.
Evidence	3: Multiple pieces of information as supporting evidence. 2: One supporting information. 1: No information as evidence.
Rebuttal	3: Multiple rebuttals targeting specific parts of the HS. 2: One rebuttal. 1: No rebuttal.
Fairness	3: Respectful, no swearing, no attacks against the hater. 2: Mostly polite, no attacks, mild offensive. 1: Aggressive, including swearing and personal attacks.
Binary Dimensions (0/1)	
Emotional Appeal	1: Language that evokes emotions (either positive or negative). 0: Neutral tone and language.
Audience Adaptation	1: Matches HS tone/complexity. 0: Mismatched HS tone or complexity.

Unlike prior frameworks, focused narrowly on toxicity or engagement, our approach enables fine-grained analysis informed by the theoretical perspectives described in argumentation and communication studies. They are divided into binary and categorical metrics.

Binary metrics. Emotional Appeal and Audience Adaptation are binary (0-1), where 1 indicates presence and 0 absence of the dimension. Due to their inherent subjectiveness and context-dependency, a binary classification (presence vs. absence) seems reasonable. Similar to prior work simplifying subjective dimensions into binary formats [45], [64]–[66], we adopt this approach to reduce ambiguity and subjectivity.

Categorical metrics. Clarity, Evidence, Rebuttal, and Fairness are categorical dimensions, using a 1–3 Likert scale, with 3 as the best score. Compared to the binary variables, these dimensions rely on more objective indicators (for example, the number of pieces of evidence), therefore, a 1-3 scale allows annotators to capture gradual differences, distinguishing between weak, moderate, and strong instances. Table I reports a summary of the definitions for the annotation of the effectiveness dimensions.

IV. DATASETS DESCRIPTION

We extend the annotation on two existing datasets: **CONAN** [15] and the **Twitter Dataset** [14]. We select CONAN as it is the first expert-curated dataset of HS/CS pairs, widely used as a benchmark. It is a multilingual (English, French, Italian) dataset centered on Islamophobia, containing 4,078 expert-annotated HS/CS pairs. Through translation and paraphrasing, it is expanded to 14,988 pairs. For our experiments, we retain only the English, non-augmented instances, resulting in 3,847 pairs. The Twitter Dataset[2] is a real-world dataset, containing 5,652 hateful tweets and replies obtained from social media (Twitter/X), capturing the brevity and style typical of online discourse. We focus on the subset labeled as counter-speech, where a clear target is identifiable, yielding 367 HS/CS pairs. Together, these datasets form a combined benchmark of 4,214 HS/CS pairs, used in our experiments.

Annotation & IAA. We define annotation guidelines to label CS effectiveness dimensions. To consolidate these guidelines[3], a pilot study was conducted on 50 HS/CS pairs from each dataset, involving three annotators [4] with background in computer science and linguistics. Through different reconciliation phases, the guidelines were improved iteratively. After an initial round, annotators reviewed disagreements to resolve confusion or misinterpretation. Based on their feedback, the guidelines were updated, and a second and final annotation round was conducted. To measure inter-annotator agreement (IAA), Fleiss's κ was used for binary dimensions, Krippendorff's α for categorical ones, and Percent Agreement for Audience Adaptation due to near-perfect consensus (as other metrics underestimate high non-random agreement). Table II shows IAA results among the three annotators. The IAA of the second round of annotations improved significantly, compared to the first one, as the annotators had the possibility to discuss and refine their understanding of the guidelines. Overall, the final IAA scores indicate strong agreement across all dimensions, so the remaining part of the dataset was labeled by one of the annotators. Table II also shows average effectiveness scores across both datasets. Expert-written counter-speech obtains an average effectiveness score of 1.62, demonstrating its higher effectiveness compared to user-written counter-speech with an average score of 1.34. CONAN scores higher in all dimensions except emotional appeal, suggesting that online users rely more on emotions than evidential reasoning or facts.

TABLE II
IAA FROM TWO ANNOTATION ROUNDS (FLEISS' κ FOR BINARY VARIABLES, AND KRIPPENDORFF'S α FOR CATEGORICAL ONES) AND AVERAGE EFFECTIVENESS SCORES.

Dimension	**CONAN**			**Twitter**		
	1st	2nd	avg.	1st	2nd	avg.
Emotional	0.30	0.65	0.29	0.40	0.62	0.45
Audience	1.00	1.00	0.99	0.99	0.99	1.00
Clarity	0.29	0.75	2.72	0.41	0.82	1.96
Evidence	0.27	0.73	1.47	0.34	0.75	1.07
Rebuttal	0.21	0.78	1.31	0.42	0.62	1.18
Fairness	0.63	0.79	2.95	0.53	0.94	2.35
Total AVG	–	–	**1.62**	–	–	**1.34**

V. EXPERIMENTAL SETTING

In this section, we describe the model configurations used in our experiments. The goal is to predict CS effectiveness scores across six dimensions. We use BERT classifiers based

[2]Only tweet IDs are publicly available due to privacy policies. We obtained the full data upon request from the authors.

[3]Complete guidelines are available at this link

[4]Two female, one other; age group: 21–30; education level: PhD students.

on `bert-base-uncased` [67] to predict the effectiveness scores for a given counter-speech response. In all configurations, the `max_length` is set to 128. The `batch_size` is 16, and the `learning_rate` is 2e-5. We run the experiments using five different seeds (0, 1, 2, 3, 42), and we average the results. To perform the experiments, we concatenate both CONAN and Twitter datasets, and then we randomly split it into train, validation, and test subsets, with a percentage of 70-10-20. We also perform cross-validation, by using a single dataset for training and validation, and the other for testing. The experiments were run on one A100 GPU.

We use BERT as the sole model to focus on improving it through our framework. By not comparing across different architectures, we isolate the effect of our proposed models, ensuring that observed gains result from our contributions, rather than differences between model architectures.

We assess the following configurations:

- **Bert_CS**: counter-speech text embeddings are used as input. Binary dimensions are trained using the Binary Cross-Entropy (BCE) loss function, and categorical ones with Cross-Entropy Loss (CE).
- **Bert_CS_HS**: identical to Bert_CS, but the input is the concatenation of hate speech and counter-speech text embeddings, to add more context.
- **Multi-task_divided.** Binary dimensions are trained jointly using BCE (for *emotional_appeal*) and Focal Loss (to mitigate class imbalance for *audience_adaptation*), while categorical dimensions are trained separately using CE. Final predictions are obtained by combining all loss components. This setting assumes partial correlation among grouped dimensions.
- **Multi-task_united.** All dimensions (binary and categorical) are trained together using their respective loss functions in a single multi-task model. The total loss is computed as the sum of individual losses. This configuration assumes stronger interdependence among all dimensions.
- **Dependency_matrix_3e.** All the dimensions are considered together, with their respective loss function (BCE for *emotional_appeal*, Focal Loss for *audience_adaptation*, and Cross Entropy Loss for all the multi-label dimensions), which are summed up at the end. Additionally, there is a new parameter, a `learnable dependency matrix`, to capture pairwise relationships between dimensions. Its rows and columns contain values between 0 and 1, that are weights accounting for the dependency between pairs of dimensions. Its weights are randomly initialized, and their best value is learned during the training phase for 3 epochs.
- **Dependency_matrix_6e** has the same configuration of Dependency_matrix_3e, but it is trained for 6 epochs to assess the effect of extended training.

We experiment with two types of input embeddings: (i) concatenated Hate Speech (HS) and Counter Speech (CS) embeddings to provide broader context and (ii) CS embeddings alone. The latter consistently yields slightly better performance. Consequently, we adopt CS-only embeddings for all our proposed configurations (*Multi-task_divided*, *Multi-task_united*, *Dependency_matrix_3e*, and *Dependency_matrix_6e*).

VI. Results

Table III shows average F1 scores across five runs for all model configurations across our counter-speech effectiveness dimensions. We compare BERT baselines, multi-task learning models, and dependency-based architectures (with CS embeddings). The dependency-based model *Dependency_matrix_6e* consistently achieves the highest F1 score (0.96), outperforming all the others. This performance gain is likely due to the use of a learnable dependency matrix that captures the interrelations between dimensions, allowing the model to exploit cross-dimensional dependencies, unlike models that treat the six dimensions independently. Statistical tests confirm that the results are significantly different. A one-way ANOVA test, conducted on the average F1 scores across models, confirms a statistically significant difference in performance ($F = 12.92$, $p < 0.001$), especially between the best model, *Dependency_matrix_6e*, and all others (pairwise t-tests between it and each of the other models show statistically significant differences at the 1% significance level).

Figure 2 shows the dependency matrix learned by *Dependency_matrix_6e*. Rows represent influenced tasks, columns represent influencing tasks, with higher values indicating stronger influence (diagonal values are zero as dimensions do not influence themselves). We can see that Emotional Appeal is influenced by Fairness (0.63) and Rebuttal (0.65), while Audience Adaptation is shaped by Clarity (0.98), Rebuttal (0.96), and Fairness (0.84). Clarity is strongly influenced by Rebuttal (0.95) and Emotional Appeal (0.83); Evidence is influenced by Audience Adaptation (0.98); Fairness is influenced by Emotional Appeal (0.80) and Audience Adaptation (0.95), and Rebuttal by Emotional Appeal (0.7). These inter-dependencies help the model to improve its classification performance across dimensions. Both multi-task models, *Multi-task_divided* (0.94) and *Multi-task_united* (0.92), also outperform the BERT baselines, suggesting that the multi-task framework is more effective, with only a slight performance difference between united and divided. Fine-tuned BERT models remain competitive: *Bert_CS* (0.91) slightly outperforms *Bert_CS_HS* (0.90). This suggests that incorporating HS data does not significantly improve classification performance.

Emotional Appeal remains the hardest dimension to classify, showing the most variation across models, ranging from 0.67 to 0.92 F1 for the combined dataset. In this dimension, dependency-based and multi-task models perform best (up to 0.92 for *Dependency_matrix_6e* and *Multi-task_divided*), while BERT baselines lag behind (0.67 and 0.62), suggesting that richer architectures better capture emotional cues. Audience Adaptation scores above 0.99 in all configurations. Although this can be due to class imbalance, we address this concern using focal loss during training, which focuses the training on hard, misclassified examples and reduces the

TABLE III

F1 SCORES (± STD) FOR ALL MODELS. STATISTICALLY SIGNIFICANT RESULTS ARE MARKED WITH *. AVG: MACRO-AVERAGE. EMO.: EMOTIONAL APPEAL, AUD.: AUDIENCE ADAPTATION, CLARITY: CLARITY, EVID.: EVIDENCE, REBUT.: REBUTTAL, FAIR.: FAIRNESS.

Model	Train: Combined Dataset							Train: Twitter → CONAN							Train: CONAN → Twitter						
	Emo.	Aud.	Clarity	Evid.	Rebut.	Fair.	AVG	Emo.	Aud.	Clarity	Evid.	Rebut.	Fair.	AVG	Emo.	Aud.	Clarity	Evid.	Rebut.	Fair.	AVG
bert_cs	$0.67^{\pm 0.09}$	$0.99^{\pm 0.00}$	$0.94^{\pm 0.01}$	$0.96^{\pm 0.01}$	$0.96^{\pm 0.01}$	$0.95^{\pm 0.01}$	$0.91^{\pm 0.02}$	$0.14^{\pm 0.00}$	$0.99^{\pm 0.00}$	$0.74^{\pm 0.05}$	$0.45^{\pm 0.06}$	$0.55^{\pm 0.11}$	$0.92^{\pm 0.01}$	$0.63^{\pm 0.03}$	$0.30^{\pm 0.01}$	$1.00^{\pm 0.00}$	$0.41^{\pm 0.04}$	$0.86^{\pm 0.09}$	$0.77^{\pm 0.02}$	$0.54^{\pm 0.21}$	$0.65^{\pm 0.05}$
bert_cs_hs	$0.62^{\pm 0.12}$	$0.99^{\pm 0.00}$	$0.94^{\pm 0.02}$	$0.96^{\pm 0.01}$	$0.95^{\pm 0.01}$	$0.94^{\pm 0.01}$	$0.90^{\pm 0.02}$	$0.14^{\pm 0.01}$	$0.99^{\pm 0.00}$	$0.76^{\pm 0.06}$	$0.43^{\pm 0.02}$	$0.55^{\pm 0.14}$	$0.92^{\pm 0.02}$	$0.63^{\pm 0.03}$	$0.30^{\pm 0.04}$	$1.00^{\pm 0.00}$	$0.33^{\pm 0.03}$	$0.84^{\pm 0.07}$	$0.73^{\pm 0.05}$	$0.41^{\pm 0.07}$	$0.60^{\pm 0.02}$
multi-task_d	$0.92^{\pm 0.02}$	$0.99^{\pm 0.00}$	$0.92^{\pm 0.01}$	$0.93^{\pm 0.01}$	$0.94^{\pm 0.03}$	$0.92^{\pm 0.03}$	$0.94^{\pm 0.01}$	$0.55^{\pm 0.17}$	$0.99^{\pm 0.00}$	$0.51^{\pm 0.24}$	$0.44^{\pm 0.01}$	$0.60^{\pm 0.05}$	$0.89^{\pm 0.13}$	**0.66*** $^{\pm 0.04}$	$0.49^{\pm 0.07}$	$1.00^{\pm 0.00}$	$0.42^{\pm 0.02}$	$0.91^{\pm 0.04}$	$0.80^{\pm 0.02}$	$0.32^{\pm 0.03}$	$0.65^{\pm 0.02}$
multi-task_u	$0.84^{\pm 0.04}$	$0.99^{\pm 0.00}$	$0.93^{\pm 0.02}$	$0.92^{\pm 0.02}$	$0.93^{\pm 0.03}$	$0.93^{\pm 0.02}$	$0.92^{\pm 0.01}$	$0.47^{\pm 0.23}$	$0.99^{\pm 0.00}$	$0.38^{\pm 0.27}$	$0.43^{\pm 0.02}$	$0.56^{\pm 0.07}$	$0.72^{\pm 0.39}$	$0.59^{\pm 0.09}$	$0.52^{\pm 0.06}$	$1.00^{\pm 0.00}$	$0.42^{\pm 0.03}$	$0.93^{\pm 0.04}$	$0.77^{\pm 0.02}$	$0.35^{\pm 0.05}$	$0.66^{\pm 0.01}$
dep._m._3e	$0.80^{\pm 0.10}$	$0.99^{\pm 0.00}$	$0.92^{\pm 0.01}$	$0.90^{\pm 0.05}$	$0.91^{\pm 0.03}$	$0.93^{\pm 0.03}$	$0.91^{\pm 0.02}$	$0.40^{\pm 0.24}$	$0.99^{\pm 0.00}$	$0.48^{\pm 0.22}$	$0.43^{\pm 0.03}$	$0.53^{\pm 0.09}$	$0.70^{\pm 0.13}$	$0.59^{\pm 0.08}$	$0.54^{\pm 0.02}$	$1.00^{\pm 0.00}$	$0.41^{\pm 0.02}$	$0.92^{\pm 0.02}$	$0.77^{\pm 0.02}$	$0.34^{\pm 0.05}$	$0.66^{\pm 0.01}$
dep._m._6e	$0.92^{\pm 0.04}$	$0.99^{\pm 0.00}$	$0.95^{\pm 0.01}$	$0.97^{\pm 0.02}$	$0.96^{\pm 0.01}$	$0.95^{\pm 0.02}$	**0.96*** $^{\pm 0.01}$	$0.40^{\pm 0.21}$	$0.99^{\pm 0.00}$	$0.75^{\pm 0.04}$	$0.53^{\pm 0.07}$	$0.47^{\pm 0.21}$	$0.78^{\pm 0.14}$	$0.65^{\pm 0.07}$	$0.50^{\pm 0.05}$	$1.00^{\pm 0.00}$	$0.42^{\pm 0.01}$	$0.92^{\pm 0.03}$	$0.77^{\pm 0.02}$	$0.38^{\pm 0.01}$	**0.66*** $^{\pm 0.01}$

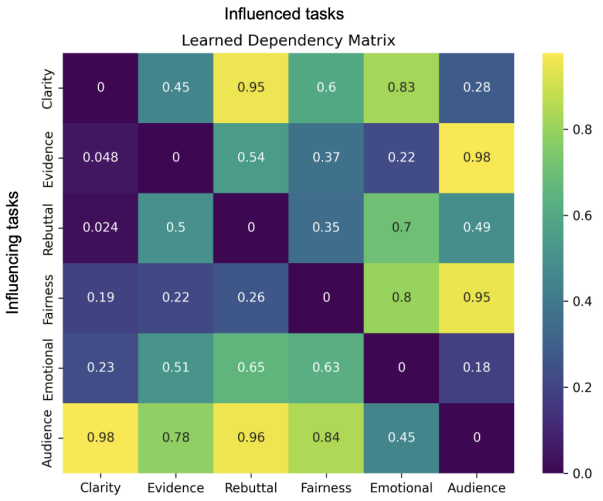

Fig. 2. Learned Dependency Matrix obtained from *Dependency_matrix_6e* on the combined dataset with seed 42.

impact of easy, well-classified examples. By contrast, Clarity and Evidence show greater variation, pointing to their higher classification difficulty.

Cross-Domain Generalization. To assess model generalization, we perform cross-domain evaluation. In the first setting, models are trained on CONAN (expert-written CS) and tested on Twitter Dataset (user-generated CS). All models show a marked performance drop, with *Dependency_matrix_6e* performing best (0.66) alongside *Dependency_matrix_3e* and *Multi-task_united*. Emotional Appeal declines the most, suggesting a domain shift in how emotional content is conveyed. Clarity, Fairness, and Rebuttal scores are also low, reflecting the less structured nature of user-written CS. In the reverse setting, where models are trained on Twitter Dataset (user-written CS) and tested on CONAN (expert-generated CS), results are similar. *Multi-task_divided* achieves the highest F1 score (0.66), followed closely by *Dependency_matrix_6e* (0.65), which remains highly competitive. The remaining models obtain lower average F1 scores, indicating less robust generalization in this transfer setup. Models trained on expert-written CS (Twitter → CONAN) seem more adaptable, but Emotional Appeal remains challenging across domains, ranging from 0.14 to 0.55. In contrast, Audience Adaptation remains consistently high, showing robustness across domains, while Fairness remains strong only on expert-written data. Overall, our models generalize reasonably well, achieving a good F1 score in these cross-dataset tests. This suggests that models learn meaningful, transferable patterns beyond dataset-specific characteristics. Given the variability in emotional appeal, domain adaptation strategies or dataset augmentation could help models better capture diverse emotional expressions across different counter-speech sources.

In-Domain Evaluation. Table IV reports F1 classification results of model configurations, with training and testing on the same dataset for the Twitter Dataset and for CONAN. Training and testing on the same dataset confirm that dependency-based models consistently perform best across effectiveness dimensions, with *Dependency_matrix_6e* again achieving the best average F1 score, further supporting the benefit of modeling interrelations between dimensions in classification tasks. We also notice that training and testing on expert-written CS (CONAN) seem easier, as it achieves higher average F1 scores, compared to training and testing on the Twitter dataset.

The results across all three evaluation settings highlight key insights into the classification of counter-speech effectiveness dimensions. While emotional appeal presents unique challenges due to its subjectivity, the other dimensions (audience adaptation, clarity, evidence, rebuttal, and fairness) also exhibit distinct patterns that provide valuable directions for model improvement.

VII. ERROR ANALYSIS AND DISCUSSION

In this section, we discuss some examples to better understand the models' behaviors on two dimensions, Emotional Appeal (most difficult to classify) and Audience Adaptation (highest F1 scores), and we address some limitations.

Audience Adaptation. It consistently achieves high F1 scores (0.99–1.00) across models, likely due to the similarity in linguistic structure and content between CS and HS. For example, consider the following HS/CS pair from CONAN:

HS: *"Islam is not a religion that professes peace and it does not make any good men. It condemns women and encourages delinquency. Lawbreakers are not all of a certain race or religion, but only one race and religion make lawbreakers."*

TABLE IV
F1 SCORES PER EFFECTIVENESS DIMENSION ON TWITTER AND CONAN DATASETS. AVG: MACRO-AVERAGE. EMO.: EMOTIONAL APPEAL, AUD.: AUDIENCE ADAPTATION, CLARITY: CLARITY, EVID.: EVIDENCE, REBUT.: REBUTTAL, FAIR.: FAIRNESS.

Model	Twitter Dataset							CONAN Dataset						
	Emo.	**Aud.**	**Clarity**	**Evid.**	**Rebut.**	**Fair.**	**AVG**	**Emo.**	**Aud.**	**Clarity**	**Evid.**	**Rebut.**	**Fair.**	**AVG**
bert_cs	0.31 ± 0.04	1.00 ± 0.00	0.52 ± 0.11	0.91 ± 0.03	0.78 ± 0.06	0.54 ± 0.09	0.68 ± 0.02	0.61 ± 0.15	0.99 ± 0.00	0.98 ± 0.01	0.97 ± 0.01	0.97 ± 0.01	0.99 ± 0.00	0.92 ± 0.03
bert_cs_hs	0.30 ± 0.04	1.00 ± 0.00	0.49 ± 0.16	0.91 ± 0.04	0.77 ± 0.07	0.48 ± 0.07	0.66 ± 0.02	0.69 ± 0.13	0.99 ± 0.00	0.97 ± 0.01	0.97 ± 0.02	0.88 ± 0.19	0.97 ± 0.03	0.91 ± 0.05
multi-task_d.	0.51 ± 0.00	1.00 ± 0.00	0.45 ± 0.00	0.97 ± 0.00	0.62 ± 0.00	0.49 ± 0.00	0.67 ± 0.00	0.95 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	0.95 ± 0.00	0.96 ± 0.00
multi-task_u.	0.47 ± 0.00	1.00 ± 0.00	0.31 ± 0.00	0.97 ± 0.00	0.60 ± 0.00	0.45 ± 0.00	0.63 ± 0.00	0.91 ± 0.00	1.00 ± 0.00	0.96 ± 0.00	0.94 ± 0.00	0.96 ± 0.00	0.96 ± 0.00	0.96 ± 0.00
dependency_m._3e	0.45 ± 0.00	1.00 ± 0.00	0.32 ± 0.00	0.97 ± 0.00	0.60 ± 0.00	0.51 ± 0.00	0.64 ± 0.00	0.86 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	0.95 ± 0.00	0.96 ± 0.00	0.98 ± 0.00	0.95 ± 0.00
dependency_m._6e	0.59 ± 0.00	1.00 ± 0.00	0.41 ± 0.00	0.97 ± 0.00	0.60 ± 0.00	0.61 ± 0.00	**0.70** ± 0.00	0.95 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00	**0.98** ± 0.00

CS: *"Today, in many Muslim-majority communities, women receive equal, even higher in some places, treatment as men. For example, many women work as scientists, engineers, lawyers, etc."* Both speeches aim to persuade a specific group by matching linguistic complexity, thus ensuring audience adaptation. The CS provides counter-examples to the HS claims, using similarly complex and detailed arguments. This makes it easier for the classifier to recognize this dimension, as both speeches maintain a similar linguistic and argumentative structure. However, this near-perfect performance could reflect data bias, suggesting that audience adaptation might be too easily detected, due to the class imbalance. Further error analysis should investigate whether models are truly understanding adaptation strategies or simply leveraging surface-level cues.

Emotional Appeal. Emotional appeal is challenging to classify due to its subjective nature and context-dependent expression. For example, consider the following counter-speeches from Twitter and CONAN:

Twitter: *"@user this isn't funny at all. don't make jokes of the situation when rene herself is affected by it. don't get her dragged even more and delete this."* This CS is direct and uses imperatives to express frustration and urgency, relying on internet culture and social cues like tagging someone and using second-person pronouns ("you").

CONAN: *"I thought we'd stopped exporting our convicted criminals last century. Now you advocate exporting people without a trial or a conviction."* This example uses sarcasm, rhetorical questioning, and historical references to evoke moral outrage and challenge the opposing argument. The emotional appeal here is more subtle and indirect, contrasting with the explicit frustration in the Twitter example. A model trained solely on Twitter might miss these subtleties found in expert-written CS. To improve emotional appeal classification, domain adaptation or dataset augmentation could help models recognize diverse emotional expressions.

Model Limitations. We use BERT-based classifiers throughout to ensure that improvements stem from our multi-task and dependency-based framework rather than the underlying model. Our goal was not to benchmark architectures but to evaluate the framework itself. For example, while Large Language Models (LLMs) are promising, and can achieve good results, they are optimized for generation rather than classification tasks. Evaluating LLMs would also require a separate experimental design, which is beyond the scope of this paper, and we leave these comparisons as future work.

VIII. CONCLUSION

As a first contribution of this paper, we release a novel linguistic resource to the community, enriching the CONAN and Twitter dataset with manual annotations of six fine-grained dimensions of counter-speech effectiveness: Emotional Appeal, Audience Adaptation, Clarity, Evidence, Rebuttal, and Fairness. This new annotation layer provides a richer, more diverse foundation for research on CS evaluation, considering content from both expert- and user-generated sources.

As a second contribution, we introduce a framework for automatically classifying counter-speech effectiveness across these dimensions. Our approach is architecture-agnostic and compatible with a range of transformer backbones. Experimental results demonstrate that a dependency-based model, which explicitly models inter-dependencies among the dimensions, significantly outperforms standard BERT baselines, achieving 0.96 F1 score on the combined dataset. This includes strong performance on both expert- (CONAN) and user-written (Twitter) CS. In addition, our multi-task learning models also surpass BERT with 0.94 F1 score, reinforcing the hypothesis that the six dimensions are interrelated and can be jointly learned more effectively. Our cross-domain evaluation reveals substantial domain shifts between CONAN and Twitter, especially in Emotional Appeal, highlighting the challenges of generalization and the need for domain-adaptive modeling techniques. Despite this, our framework demonstrates robust improvements over BERT baselines across domains, suggesting its effectiveness can generalize beyond specific model architectures.

Finally, we conduct a detailed linguistic analysis of model performance across dimensions, revealing which aspects of CS effectiveness remain the most difficult to capture. This analysis can inform the design of future CS evaluation systems and guide annotation efforts where additional human oversight is most valuable.

ACKNOWLEDGMENT

This work has been partially supported by the French government, through the 3IA Côte d'Azur investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

REFERENCES

[1] C. Schieb and M. Preuss, "Governing hate speech by means of counter-speech on facebook," in *66th ICA Annual Conference*, Fukuoka, Japan, 2016, pp. 1–23.

[2] S. Benesch, "Countering dangerous speech: New ideas for genocide prevention," Washington, DC, 2014.

[3] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherjee, "Thou shalt not hate: Countering online hate speech," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 369–380.

[4] M. Ashida and M. Komachi, "Towards automatic generation of messages countering online hate speech and microaggressions," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, and Z. Talat, Eds. Seattle, Washington (Hybrid): Association for Computational Linguistics, Jul. 2022, pp. 11–23.

[5] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, and R. Tromble, "Introducing cad: The contextual abuse dataset," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2289–2303.

[6] D. Hangartner, G. Gennaro, S. Alasiri, N. Bahrich, A. Bornhoft, J. Boucher, B. B. Demirci, L. Derksen, A. Hall, M. Jochum *et al.*, "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment," *Proceedings of the National Academy of Sciences*, vol. 118, no. 50, p. e2116310118, 2021.

[7] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, and R. Tromble, "Detecting east asian prejudice on social media," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 2020, pp. 162–172.

[8] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar, "Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 90–94.

[9] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4755–4764.

[10] D. Rieger, J. B. Schmitt, and L. Frischlich, "Hate and counter-voices in the internet: Introduction to the special issue," *SCM Studies in Communication and Media*, vol. 7, no. 4, pp. 459–472, 2018.

[11] J. Mun, E. Allaway, A. Yerukola, L. Vianna, S.-J. Leslie, and M. Sap, "Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9759–9777.

[12] S. S. Tekiroğlu, H. Bonaldi, M. Fanton, and M. Guerini, "Using pre-trained language models for producing counter narratives against hate speech: a comparative study," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3099–3114.

[13] S. S. Tekiroğlu, Y.-L. Chung, and M. Guerini, "Generating counter narratives against online hate speech: Data and strategies," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1177–1190.

[14] A. Albanyan and E. Blanco, "Pinpointing fine-grained relationships between hateful tweets and replies," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 10 418–10 426.

[15] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2819–2829.

[16] Y.-L. Chung, S. S. Tekiroğlu, and M. Guerini, "Towards knowledge-grounded counter narrative generation for hate speech," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 899–914.

[17] W. Zhu and S. Bhat, "Generate, prune, select: A pipeline for counterspeech generation against online hate speech," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 134–149.

[18] A. Baheti, M. Sap, A. Ritter, and M. Riedl, "Just say no: Analyzing the stance of neural dialogue generation in offensive contexts," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4846–4862.

[19] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[20] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[21] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts, "An expert annotated dataset for the detection of online misogyny," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 1336–1350.

[22] L. Grimminger and R. Klinger, "Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2021, pp. 171–180.

[23] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," *CoRR*, vol. abs/1701.08118, 2017. [Online]. Available: https://arxiv.org/abs/1701.08118

[24] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, "Social bias frames: Reasoning about social and power implications of language," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 5477–5490.

[25] A. Hede, O. Agarwal, L. Lu, D. C. Mutz, and A. Nenkova, "From toxicity in online comments to incivility in american news: Proceed with caution," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 2620–2630.

[26] M. Wiegand, J. Ruppenhofer, and E. Eder, "Implicitly abusive language – what does it actually look like and why are we not getting there?" in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021, pp. 576–587.

[27] Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, 2016, pp. 138–142.

[28] N. A. A. Aziz, M. A. Maarof, and A. Zainal, "Hate speech and offensive language detection: A new feature set with filter-embedded combining feature selection," in *2021 3rd International Cyber Resilience Conference (CRC)*. IEEE, 2021, pp. 1–6.

[29] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: Experimenting with hindi and english social media," *Information*, vol. 12, no. 1, p. 5, 2021.

[30] S. Khan, M. Fazil, A. L. Imoize, B. I. Alabduallah, B. M. Albahlal, S. A. Alajlan, A. Almjally, and T. Siddiqui, "Transformer architecture-based transfer learning for politeness prediction in conversation," *Sustainability*, vol. 15, no. 14, p. 10828, 2023.

[31] M. A. H. Wadud, M. Mridha, J. Shin, K. Nur, and A. K. Saha, "Deep-bert: Transfer learning for classifying multilingual offensive texts on social media," *Computer Systems Science & Engineering*, vol. 44, no. 2, 2023.

[32] J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos, "Deeper attention to abusive user content moderation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[33] B. Susan, R. Derek, P. D. Kelly, S. M. Haji, and W. Lucas, "Counter-speech on twitter: A field study," 2016.

[34] J. Garland, K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, and M. Galesic, "Countering hate on social media: Large scale classification of hate and counter speech," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, S. Akiwowo, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 102–112.

[35] A. Poudhar, I. Konstas, and G. Abercrombie, "A strategy labelled dataset of counterspeech," in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, and A. Calabrese, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 256–265.

[36] B. Mathew, N. Kumar, P. Goyal, and A. Mukherjee, "Analyzing the hate and counter speech accounts on twitter," *arXiv*, 2018.

[37] L. Wright, D. Ruths, K. P. Dillon, H. M. Saleem, and S. Benesch, "Vectors for counterspeech on twitter," in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 57–62.

[38] C. Ziems, B. He, S. Soni, and S. Kumar, "Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis," *arXiv*, 2020.

[39] Y. Zheng, B. Ross, and W. Magdy, "What makes good counterspeech? a comparison of generation approaches and evaluation metrics," in *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, Y.-L. Chung, H. Bonaldi, G. Abercrombie, and M. Guerini, Eds. Prague, Czechia: Association for Computational Linguistics, Sep. 2023, pp. 62–71.

[40] J. Jones, L. Mo, E. Fosler-Lussier, and H. Sun, "A multi-aspect framework for counter narrative evaluation using large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 147–168.

[41] X. Yu, E. Blanco, and L. Hong, "Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech," in *Proceedings of the 18th International AAAI Conference on Web and Social Media (ICWSM)*, 2024, pp. 1740–1752.

[42] X. Song, S. Mamidisetty, E. Blanco, and L. Hong, "Assessing the human likeness of AI-generated counterspeech," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 3547–3559.

[43] I. Persing and V. Ng, "Modeling thesis clarity in student essays," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, H. Schuetze, P. Fung, and M. Poesio, Eds. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 260–269.

[44] H. Wachsmuth, K. Al-Khatib, and B. Stein, "Using argument mining to assess the argumentation quality of essays," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Y. Matsumoto and R. Prasad, Eds. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1680–1691.

[45] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. A. Thijm, G. Hirst, and B. Stein, "Computational argumentation quality assessment in natural language," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187.

[46] P. Stapleton and Y. Wu, "Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance," *Journal of English for Academic Purposes*, vol. 17, pp. 12–23, 2015.

[47] R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, and N. Slonim, "Show me your evidence - an automatic method for context dependent evidence detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 440–450.

[48] Z. Rahimi, D. J. Litman, R. Correnti, L. C. Matsumura, E. Wang, and Z. Kisa, "Automatic scoring of an analytical response-to-text assessment," in *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, 2014, pp. 601–610.

[49] J. C. Reinard, "The empirical study of the persuasive effects of evidence: The status after fifty years of research," *Human Communication Research*, vol. 15, no. 1, pp. 3–59, 1988.

[50] Aristotle, *On Rhetoric: A Theory of Civic Discourse*, ser. Clarendon Aristotle Series. Oxford University Press, 2007.

[51] T. Govier, *A Practical Study of Argument*, 7th ed. Belmont, CA: Wadsworth, Cengage Learning, 2010.

[52] J. Chen, Y. Yan, and J. Leach, "Are emotion-expressing messages more shared on social media? a meta-analytic review," *Review of Communication Research*, vol. N/A, p. 21, 2022.

[53] V. Marone, "Online humour as a community-building cushioning glue," *The European Journal of Humour Research*, vol. 3, no. 1, pp. 61–83, 2015.

[54] S. E. Toulmin, *The Uses of Argument*. Cambridge University Press, 1958.

[55] T. E. Damer, *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*, 6th ed. Belmont, CA: Wadsworth, Cengage Learning, 2009.

[56] S. Granger, E. Dagneaux, F. Meunier, and M. Paquot, *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain, 2009.

[57] D. J. O'Keefe, "How to handle opposing arguments in persuasive messages: A meta-analytic review of the effects of one-sided and two-sided messages," *Annals of the International Communication Association*, vol. 22, no. 1, pp. 209–249, 1999.

[58] R. Onoda, S. Miwa, and K. Akita, "Highlighting effect: The function of rebuttals in written argument." in *EAPCogSci*, 2015.

[59] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[60] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words!: linguistic style accommodation in social media," in *Proceedings of the 20th International Conference on World Wide Web*. ACM, 2011, pp. 745–754.

[61] Y. Wang, D. Reitter, and J. Yen, "A model to qualify the linguistic adaptation phenomenon in online conversation threads: Analyzing priming effect in online health community," in *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, V. Demberg and T. O'Donnell, Eds. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 55–62.

[62] R. B. Cialdini, *Influence: Science and Practice, rev. ed.* New York: William Morrow, 1993.

[63] M. Schreier, N. Groeben, and U. Christmann, "That's not fair! argumentational integrity as an ethics of argumentative communication," *Argumentation*, vol. 9, no. 2, pp. 267–289, 1995.

[64] I. Habernal and I. Gurevych, "Which argument is more convincing? analyzing and predicting convincingness across topics," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1589–1599.

[65] S. Oraby, L. Reed, and M. Walker, "And that's a fact: Distinguishing factual and emotional argumentation in online dialogue," in *Proceedings of the 4th Workshop on Argument Mining*, 2017.

[66] M. Koszowy, P. Bosc *et al.*, "Pathos in natural language argumentation: Emotional appeals and reactions," *Argumentation*, 2024.

[67] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.