Old but Gold: LLM-Based Features and Shallow Learning Methods for Fine-Grained Controversy Analysis in YouTube Comments

Davide Bassi[1], **Erik Bran Marino**[2], **Renata Vieira**[2], **Martin Pereira-Farina**[3]
[1] CiTIUS, University of Santiago de Compostela, Spain, davide.bassi@usc.es
[2] CIDEHUS, University of Évora, Portugal, {erik.marino,renatav}@uevora.pt
[3] University of Santiago de Compostela, Spain, martin.pereira@usc.es

Abstract

Online discussions can either bridge differences through constructive dialogue or amplify divisions through destructive interactions. This paper proposes a computational approach to analyze dialogical relation patterns in YouTube comments, offering a fine-grained framework for controversy detection, enabling also analysis of individual contributions. Our experiments demonstrate that shallow learning methods, when equipped with theoretically-grounded features, consistently outperform more complex language models in characterizing discourse quality at both comment-pair and conversation-chain levels. Ablation studies confirm that divisive rhetorical techniques serve as strong predictors of destructive communication patterns. This work advances understanding of how communicative choices shape online discourse, moving beyond engagement metrics toward nuanced examination of constructive versus destructive dialogue patterns.

1 Introduction

Online discussions can either bridge differences through constructive dialogue or amplify divisions through inflammatory responses. These divergent outcomes are fundamentally shaped by the communicative approaches adopted by participants, where each contribution can either push the interaction toward controversy or constructive discussion. In fact, while conflicting viewpoints form a prerequisite for argumentation (Walton, 2008), individuals express opposition through diverse communicative approaches, generating a "disagreement space" that participants navigate based on their chosen discursive strategies (Schumann and Oswald, 2024).

Delineating this conceptual space is essential for advancing argument mining research, as it provides a structured framework for analyzing how disagreements manifest in discourse, enabling more nuanced computational modeling of argumentative interactions in both online and offline contexts. Samson and Nowak (2010) proposes a framework in which constructive and destructive conflicts are opposite ends of a single dimension (Vallacher et al., 2013). Specifically, destructive processes aim at inflicting psychological, material or physical damage on the opponent, while constructive aim at achieving one's goals while maintaining or enhancing relations with the opponent.

Computational approaches to detect and measure constructive versus destructive dialogical relations patterns remain underdeveloped (Lawrence and Reed, 2019). Research in this direction could generate methods useful not only to identify controversies but also to track how they emerge and evolve through specific communicative choices, advancing our understanding of these dynamics while offering practical applications for fostering healthier online discourse (Marres, 2015).

This research proposes an automated approach to classify and measure destructive and constructive patterns in online discussions, examining how individual messages, situated within their conversational context, contribute to either productive dialogue or increased antagonism.

Specifically, we contribute by: (*i*) providing a pipeline to mine laypeople discussions from Youtube video comments section and creating a dataset of full conversation chains with varied length and complexity[1]; (*ii*) proposing a novel operationalization of destructive communication through divisive rhetorical techniques (Zompetti, 2015), demonstrating how stance, linguistic and rhetorical features can be used to characterize dialogical quality in online discourse; (*iii*) evaluating the effectiveness of these rhetorical features across both traditional machine learning methods and large language models, with results revealing

[1]Full dataset, annotation guidelines and the scripts we used can be found at https://github.com/BassiDavide/Arg-Mining_Old_but_Gold/tree/main.

Proceedings of the 12th Argument Mining Workshop, pages 46–57

that shallow learning approaches more effectively leverage them for classification.

2 Related Works

Our research advances the field of controversy analysis, diverging from predominant approaches reliant on quantitative engagement metrics and network-based methodologies (Coletto et al., 2017; Sriteja et al., 2017; Garimella et al., 2016). Instead, we employ a finer-grained, textually grounded framework, akin to Wang et al. (2023); Konat et al. (2016); Allen et al. (2014); Chen et al. (2023), to dissect discursive comment-level interactions and derive higher-level insights about conversation quality, i.e. distinguishing between destructive (controversy-promoting) and constructive communication patterns at both comment and comment-chain levels. Additionally, our study represents the first controversy analysis of YouTube discussions—a platform that, despite its ubiquity, remains understudied through this analytical lens because of its API limitations. To tackle these issues we employ Bassi et al. (2024b)'s pipeline to extract conversation structures and stance information.

Prior efforts aimed to tackle the multifaceted nature of evaluating dialogue quality, yielding valuable insights. Samson and Nowak (2010) established that constructive and destructive conflict processes can be distinguished through linguistic markers (e.g., pronoun usage, emotional valence). Similarly, Chen et al. (2023) found that controversial comments tend to express higher levels of emotions. De Kock and Vlachos (2021) specifically investigate constructive disagreement in Wikipedia Talk pages, demonstrating that gradient features capturing temporal changes in linguistic markers and conversation structure information effectively predict escalation to mediation as a proxy for discourse failure.

Further, Lawrence et al. (2017) and Harris et al. (2018) demonstrate the significance of rhetorical figures in discursive relation detection (see also Lawrence and Reed (2019)).

To provide a more comprehensive understanding of dialogue dynamics, our approach integrates linguistic indicators with rhetorical formally detectable patterns. Specifically, we leverage Zompetti (2015)'s divisive rhetoric framework, defining specific rhetorical devices and argumentative fallacies that systematically undermine constructive dialogue.

Through this comprehensive set of features (linguistic, stance, and rhetorical devices), we develop a computational method that operates at both comment-pair and conversation-chain levels to assess discourse quality. At the micro level, we classify the relationship between adjacent comments according to their functional orientation and communicative quality. At the macro level, we aggregate these classifications to characterize entire conversation chains on a divisiveness scale from highly destructive to constructive. This multi-level approach quantifies how individual interactions contribute to broader conversational dynamics, revealing patterns that either foster productive dialogue or amplify division throughout extended discussions.

The remainder of this paper is structured as follows: Section 3 details our dataset creation and annotation. Section 4 describes features for divisiveness detection. Section 5 outlines our experimental methodology. Section 6 presents results and feature importance analysis. Section 7 discusses result implications and future work.

3 Dataset

3.1 Dataset Creation

A) Data Crawling: given our focus on controversial topics, we centered our investigation on immigration-related content. To gather the data, first, we crawled YouTube to identify the 100 most viewed videos using query sets designed to capture diverse viewpoints (see repository for complete query). We restricted our sample to English-language content from the United States (2013-2024) with a minimum threshold of 1,000 comments per video. These videos were then ranked based on their comment volume to identify those generating more discussions. We took the 15 most commented ones.

B) Conversation Reconstruction: to reconstruct conversational structures, we applied the methodology proposed by Bassi et al. (2024b), which allows to address complex dialogical discourse phenomena where the meaning of a locution can only be understood by reference to another e.g. "Isn't illegal immigration a crime?" — "Definitely not".

C) Discussion Chain Extraction: we define a discussion chain as a sequence of interconnected messages that form a coherent conversation thread. To identify and extract them, firstly, we identified *terminal messages*, i.e. messages that (*i*) have not

received no further responses; (*ii*) have a depth level of at least 4 in the conversation tree, ensuring a minimum of 5 messages in the conversation (see dotted comments in Figure 1). Otherwise, the chain was not considered (see Case-A in Figure 1). Secondly, for each identified terminal message, we *traced back through the conversation tree to the root message*, creating a complete discussion chain.

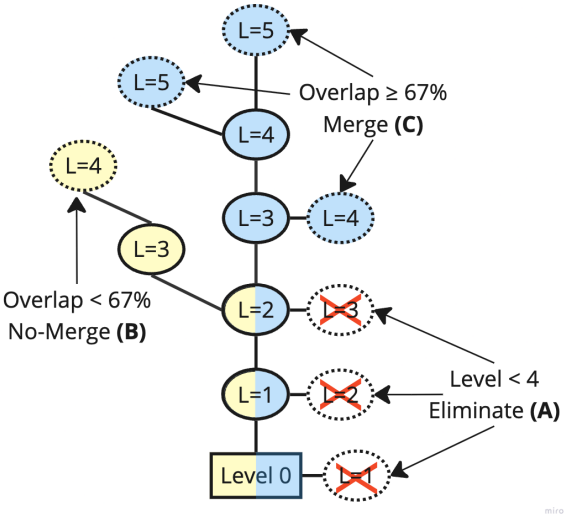


Figure 1: Chain Clustering and Construction Criteria

D) Chain Clustering and Refinement: conversation chains can present a "natural clustering" (e.g. Chain A Figure 6 in Appendix), or share some messages to, then, develop specific paths (e.g. Chain B-C-D-E in Figure 6 in Appendix). Rather than treating entire conversation trees as monolithic units, we aimed to identify and analyze these distinct conversation branches separately, as they often exhibit unique communicative patterns and divisiveness characteristics, even when originating from the same root comment. Given this "behavior", we aimed at treating these cases as separated conversations (see how we split Chain C-D-E in Figure 6 in Appendix, despite sharing messages). To systematically operate this clustering, we analyzed the overlap between 2 conversation chains implementing a pairwise comparison of discussion chains using a similarity metric. Specifically:

(i) we calculated the intersection of messages between each pair of chains (i.e., given a couple of conversation chain with different lengths ($len(com1)$,$len(com2)$), we counted how many comments they shared = $len(intersection)$).

(ii) we calculated the overlap ratio based on the size of the intersection relative to the shorter chain as: $OverRat = \frac{len(intersection)}{\min[(len(com1),len(com2)]}$

(iii) we established a threshold of 0.67: two chains were merged if they shared more than 67% of their comments (relative to the shorter chain), and vice versa for lower values of overlap.

Computationally, we implemented this process by representing each conversation chain as a node in a graph, connecting nodes that exceed our overlap threshold (0.67), and extracting connected components to identify chains forming a cluster that should be merged[2].

To ensure the robustness of our methodology, we manually verified the accuracy of comment links during annotation, identifying only 30 incorrect links out of 2387 total child-comments, confirming the reliability of our approach.

E) Sampling: we considered that conversation chains can have different degrees of complexity. To ensure a representative sample across all complexity levels, we implemented a stratified sampling approach. First, we grouped chains sharing the same root comment (Level=0) into "chain families" to preserve the contextual integrity of discussions. Each family's complexity was measured by its total message count. We then divided these families into three equal percentile groups (low, medium, and high complexity) and sampled proportionally from each group to reach our target message count (=2500).

3.2 Annotation

Our annotation schema focuses on interactional dynamics between comment pairs. The schema evolved through expert analysis, ultimately yielding a five-category taxonomy that assigns numerical values expressing each comment's contribution toward cohesion (+) or division (-). As shown in Figure 2, this framework captures two dimensions: (1) functional relationship (agreement, disagreement, neutral) and (2) communication style (constructive versus destructive), recognizing that comments with similar positions may contribute differently to discussion quality.

The five categories are (Figure 3 depicts their relative frequencies):

Constructive Disagreement (+1): expressing disagreement while maintaining conditions for mutual

[2]Chains are considered part of the same cluster if they are connected either directly through high overlap, or indirectly through a chain of high-overlap connections. For instance, given 3 conversation chains A,B,C, where A overlap 70 with B, B overlap 70 with C, and A overlap 40 with C, A and C would still be connected by virtue of B. This is called an indirect link.

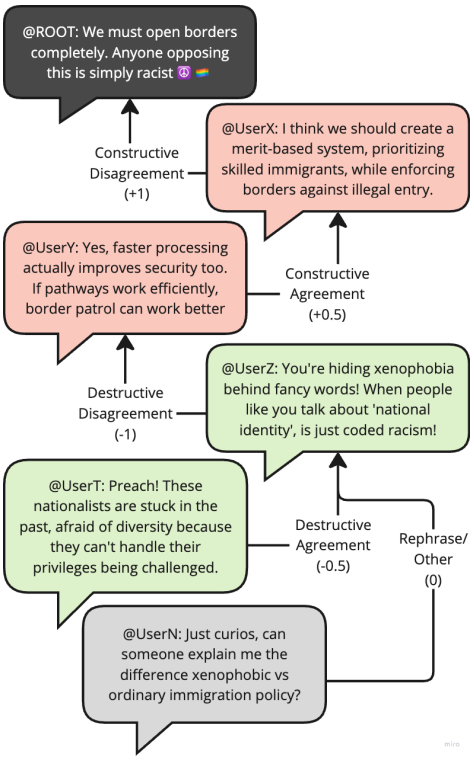


Figure 2: Annotation Example with Contra Immigration, Pro Immigration and Neutral stances interacting among each other. Arrows between messages indicate the quality of interaction.

understanding ($N = 561$)
Constructive Agreement (+0.5): strengthening mutual understanding while agreeing ($N = 203$)
Rephrasing/Neutral[3] (0): facilitating conversation without taking a stance ($N = 251$)
Destructive Agreement (-0.5): strengthening divisions while agreeing ($N = 304$)
Destructive Disagreement (-1): hindering productive dialogue through hostile language ($N = 1068$)

Two annotators were instructed to label comment pairs according to the guidelines (see repository), tracking the relation from child comment to parent comment. The messages were presented to annotators following the chronological order of the discussion, enabling them to understand the contextual flow of the conversation. Inter-rater agreement resulted in Cohen's $K = 0.37$, which, while considered fair (Landis and Koch, 1977), underscores the difficulty of operationalizing theoretical constructs of constructive versus destructive communication patterns, especially in informal online discourses.

[3]We merged neutral and rephrase, as they rarely exhibit strong constructive or destructive characteristics that would warrant separate classification.

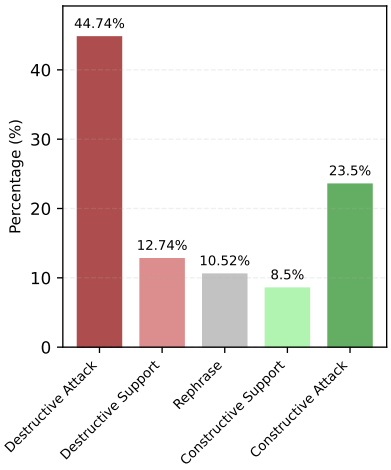


Figure 3: Relative **Comment** Labels Distribution

3.3 Conversation Chain Characterization

Category	Score Range	Count
Highly Destructive	$-1 \leq \text{HD} < -0.75$	n=33
Moderately Destructive	$-0.75 \leq \text{MD} < -0.25$	n=108
Slight/Neutral	$-0.25 \leq \text{SD} < 0.25$	n=54
Constructive	$0.25 \leq \text{C} \leq 1$	n=45

Table 1: Chain Controversy Categories Score Ranges and Distribution

The chain divisiveness categories were developed to analyze conversation chains by averaging the divisiveness values of the comments it contains. Given the strong imbalance of our messages towards the destructive side of the continuum (see Figure 4), we grouped the chain controversy scores to balance theoretical value with the empirical distribution, as shown in Table 1.

4 Features for Divisiveness Detection

4.1 Linguistic

Our analysis incorporates a diverse set of linguistic features extracted from comment text to capture communicative patterns relevant to divisiveness detection. Following Samson and Nowak (2010), for each comment, we extract linguistic elements including word count, capitals ratio, and punctuation frequencies (question and exclamation marks). We leverage VADER (Hutto and Gilbert, 2014) to obtain sentiment polarity scores (negative, positive, neutral, and compound) that capture the emotional tone of comments. Additionally, we compute parent-child comparison features to measure conversational dynamics, including word count differences and word count ratios between comments and

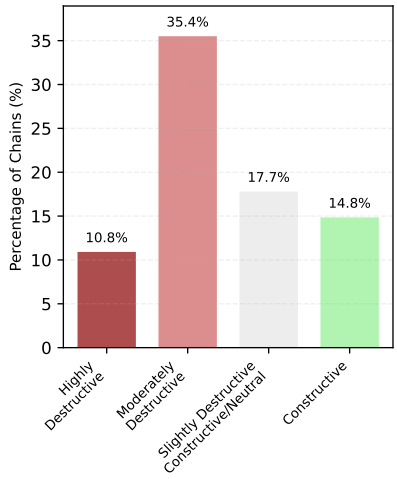


Figure 4: Relative **Chain** Distribution by Divisiveness Categories

their parents. These features aim at capturing linguistic markers of constructive versus destructive communication patterns and constitute the *Base* experimental condition in Table 2-4.

4.2 Stance

We characterize the stance of the comment towards immigration as contra, neutral, or pro using the context-sensitive approach introduced by Bassi et al. (2024b). This method leverages parent-child comment relationships to improve classification accuracy, incorporating the parent comment's stance as contextual information during classification. Following the approach of Bassi et al. (2024b), we use as classifier GPT-4o (prompt detailed in the repository). We tested the performance of the model on a manually annotated gold dataset of 1.3k comments, obtaining substantial results (macro-F1=74.5, see Table 6 in Appendix for details), which we considered robust enough to scale the method to the rest of our dataset. From these classifications, we derived *stance* and relational features such as *binary indicators for the same stance* between comment pairs, capturing the social positioning dynamics and interactions related to the topic.

4.3 Divisive Rhetorical Techniques

As outlined in Section 2, to capture the characteristic argumentative patterns of divisive discourse, we aimed at tracking a set of divisive rhetorical techniques commonly used in controversial discussions, following the work of Zompetti (2015).

Although automated detection of rhetorical techniques has traditionally employed shallow learning and encoder-based methods (Bassi et al., 2024a), Jose and Greenstadt (2024) and Sprenkamp et al. (2023) demonstrated consistent performance of LLMs without specialized training. Drawing from this, we devised a multi-label classification approach implemented through Gpt-4o-mini to identify 13 distinct techniques. The prompt provides definitions and examples for each technique to guide the classification (complete prompts can be found in the repository).

Human verification of a sample yielded SOTA-consistent performance (macro-F1=69.6, details in Table 7), allowing us to confidently apply this method to our complete dataset.

Thanks to this additional information, we generated features that quantify both the presence and frequency of these techniques in each comment, creating *binary indicators for individual techniques* and aggregate metrics like *technique count* and *binary indicator of the presence of each one*.

4.4 Embeddings

We employed SentenceTransformer (Reimers and Gurevych, 2019) to capture semantic content beyond surface features, using the "*roberta-base-nli-stsb-mean-tokens*" model to generate 768-dimensional representations. This approach offered better control over embedding dimensionality than direct BERT-based implementation, enabling more transparent integration with our theoretical features. We applied PCA during training to preserve 95% variance while reducing dimensionality, balancing semantic richness with computational efficiency; which is particularly important when combining embeddings with other feature types in shallow learning models.

5 Experiments

We operated our experiments at two distinct analytical levels: comment and chain. The first task focused on classifying the communicative relationship between parent-child comment pairs according to the five-class taxonomy described in Section 3.2. The second task evaluated how effectively comment-level predictions could characterize the overall quality of conversation chains. We aggregated individual comment scores to compute chain-level divisiveness metrics, mapping each chain to one of the categories defined in Table 1. This approach allowed us to assess the propagation of communicative patterns throughout extended conver-

sations and determine whether localized comment predictions effectively capture broader conversational dynamics.

5.1 Shallow Learning Classifiers

We evaluated several classifiers (Random Forest, Logistic Regression, SVM, and XGBoost) testing multiple combinations of features (see Section 4 and Section 6 for details) to observe the impact of each one on the performance. To address the significant class imbalance shown in Figure 3, we incorporated Synthetic Minority Over-sampling Technique (SMOTE) into our pipeline, testing different k-nearest neighbor values to find the optimal balance to avoiding overfitting on minority classes. We optimized hyperparameters through grid search with 3-fold stratified cross-validation, maximizing macro F1 scores. As detailed in Section 4.4, for embedding-rich feature sets, we applied PCA retaining 95% variance to reduce dimensionality before classification[4]. For chain-level analysis, we used predictions from the best-performing comment-level model to calculate aggregate scores, evaluating both numerical accuracy and categorical classification performance across different chain complexities.

5.2 Large Language Models

We evaluated four leading Large Language Models (LLMs) for the comment classification task: GPT-4o-mini (temp=0.1), GPT-o3-mini (effort=*medium*), DeepSeek-V3-chat (temp=0.1), and DeepSeek-R1-reasoning (temp=*not supported*). For incorporating features into the LLM approach, we designed specialized prompts for each experimental condition. The base condition used only comment text, while additional features were systematically incorporated through explicit prompt engineering: stance information was provided as categorical labels (pro/contra/neutral), rhetorical techniques were presented as a structured list with definitions, and the combined feature condition integrated all information into a single comprehensive prompt. For reasoning-enabled models, we provided explicit instructions to analyze comment relations step-by-step before determining the final classification. We created eight distinct prompts: four tailored for chat models and four designed for reasoning models. Each set of four prompts corresponded to our experimental conditions: comment text-only (*Base* condition in Table 2-4), text with stance, text with rhetorical techniques, and all features combined[5].

[4]See training scripts in repository for details.

6 Results

6.1 Comment Level

Table 2 reveals key trends in our findings. *(1)* Shallow learning models consistently outperform LLMs. *(2)* Notably, optimal performance was achieved by all shallow learning models when utilizing the comprehensive set of features, suggesting effective operationalization of constructive/destructive process concepts. To assess performance reliability, we used bootstrap resampling (1000 iterations) for LLMs and cross-validation variance for shallow learning models. Both yielded $SD \approx 0.02$, with LLM results showing tighter variance distributions than shallow learning models. Paired t-tests on key comparisons confirmed statistical significance: XGBoost (B+S+T+E) vs. DeepSeek (Base), feature engineering impact within XGBoost (Base vs. B+S+T+E), and aggregate shallow learning performance vs. LLM performance across all conditions (all $p < 0.001$). Complete bootstrap statistics are available in our repository.

Table 3 presents class-specific performance metrics for the top-performing model, revealing a degradation in model efficacy attributable to both destructive and constructive agreement classes (a trend consistently observed across all models, as detailed in Figure 7 in Appendix). This discrepancy must be contextualized within the constraints of moderate inter-annotator reliability, and the class imbalance within the dataset, which, despite the application of SMOTE, may not fully alleviate the negative impacts on model performance. Concurrently, the presence of linguistically complex phenomena, such as irony or euphemisms, as illustrated in the following example:

> Message 1 - Genuine Praise: "*Your immigration policy proposal balances security and compassion brilliantly. You really are a genius.*"
> Message 2 - Mockery/Sarcasm: "*Your solution to immigration is 'just close the borders'? You really are a genius.*"

LLMs exhibited similar difficulties with agreement categories across all models [6]. Detailed analysis reveals that the primary challenge lies not in

[5]The complete prompts can be found in the repository.

[6]Complete fine-grained performance reports are available in the repository.

Model	Base	B+Stance	B+Tech	B+S+T	B+Emb	B+S+T+E
DeepSeek	**55.11**	44.11	53.82	49.77	-	-
DeepSeek-R1	41.81	37.62	44.58	43.59	-	-
OpenAI 4o-mini	45.14	42.09	42.79	43.97	-	-
OpenAI o3-mini	44.86	43.80	46.33	46.68	-	-
Logistic Regression	28.37	39.87	45.44	55.15	38.48	57.82
SVM	28.48	42.04	45.27	51.99	39.38	58.38
XGBoost	31.87	44.67	47.81	59.24	40.27	**60.60***
Random Forest	32.23	44.30	50.55	56.78	42.42	60.14

*Best model overall for comment-level prediction
We derived Standard Deviation bootstrapping through 10000 iterations ($SD \approx 0.02$ for all)

Table 2: Comment Divisiveness Detection Performance Comparison (F1 Macro Scores)

Label	P	R	F1
Des. Disagreement (-1.0)	0.83	0.82	0.82
Des. Agreement (-0.5)	0.47	0.58	0.52
Rephrase (0.0)	0.64	0.53	0.58
Cons. Agreement (0.5)	0.41	0.58	0.48
Cons. Disagreement (1.0)	0.69	0.57	0.62
Macro Avg.	0.61	0.62	0.61
Weighted Avg.	0.69	0.68	0.68

Model: XGBoost on All Features,
PCA 40, l.rate: 0.2, max_depth: 3, n_est: 50

Table 3: Comment-Level: Class-Specific Performance

distinguishing constructive from destructive discourse, but in accurately identifying agreement versus disagreement relationships. While the minimal temperature settings (0.1) employed for chat models precluded examination of their decision-making processes—as they output only categorical labels—we report representative examples of the reasoning traces generated by reasoning-enabled models (GPT-03-mini) to illustrate potential explanations for these performance limitations.

"Parent": "Please learn about World History. [Pro]"

"Comment": "CLOSE MILITARY BASES IN 80 PLUS NATIONS CLOSE USA MILITARY BASES IN 80 PLUS NATIONS [Pro][Repetition; Slogan][7]"

"reasoning": "Step 1: The child's comment does not engage with the parent's specific idea but instead presents its own sloganized message to propose closing military bases. This does little to build mutual understanding and seems to dismiss or sidestep the parent's point. Step 2: The use of slogans and repetition indicates a tendency to oversimplify complex issues, reducing space for dialogue and critical engagement. These rhetorical techniques are characteristic of divisive messaging.Step 3: The comment lacks perspective-taking or any respectful engagement with the original idea; rather, it uses incendiary rhetoric, which is likely to escalate division. Final Score: Destructive_Disagreement/Attack",

This example demonstrates two key factors underlying LLM performance limitations. First, the model effectively utilizes the provided feature information, correctly employ the presence of divisive rhetorical techniques to justify the "Destructive Dimension" classification. However, the model fails to connect the comments to the broader immigration discussion context. Human annotators labeled this as "agreement" because both comments share an underlying anti-colonial narrative—the parent criticizing historical ignorance and the child advocating military withdrawal from former colonial territories. Hence, a key issue appears to be context-dependent agreement detection.

"Parent": "In that case maybe we should also have restrictions on smaller men, like lets say 5'7 or under since they will also have a difficult times carryinf heavy rocks?"[Neutral]

"Comment": "I thought for a moment, then I remembered in many north european contries the average is 6 and something so 5'7 is short there lol"[Neutral][No Techniques]

"reasoning": "Step 1. The child comment directly engages with the parent's observation about height and restrictions using a factual clarification while referencing the original comment. Step 2. The tone is neutral and shows a respectful, though subtly differing, perspective; it's not dismissive or mocking. Step 3. The comment contributes knowledge that can help refine the point made by the parent without attacking the person, promoting mutual understanding. Step 4. Overall, the child comment reflects a constructive disagreement. Constructive_Disagreement/Attack",

Also in this second example the model is effective in detecting the correct "tone" of the conversation. However, it misinterprets the dialogical function of the refinement of the parent's statement as an attack, when the child is actually contributing supportive evidence about height differences across countries. Hence, LLMs may struggle to distinguish between challenging refinements and supportive contributions in agreement detection. Notably, while both examples were drawn from the "Text+Stance+Techniques" condition, neither reasoning trace referenced this additional feature information, though stance features were utilized in other cases, indicating inconsistent integration. Improved prompting strategies or reasoning approaches might enhance performance. We provide complete reasoning model outputs in the repository to facilitate further error analysis.

A closer examination of LLM behavior yields interesting insights. Within this paradigm, the two sub-groups exhibit divergent responses to varying levels of feature input. *(3)* Non-reasoning ('chat') models returned their highest performance on tasks when provided with minimal feature input, with DeepSeek-V3 attaining peak performance within this category (macro-F1 = 55.11). In stark contrast, reasoning-enabled LLMs demonstrated improved performance when equipped with an extensive range of features, underscoring their enhanced capacity for leveraging supplementary information.

This disparity reflects different model design ob-

[7] See repository for complete prompt format. It included also a brief definition of the detected technique and 4 guiding questions.

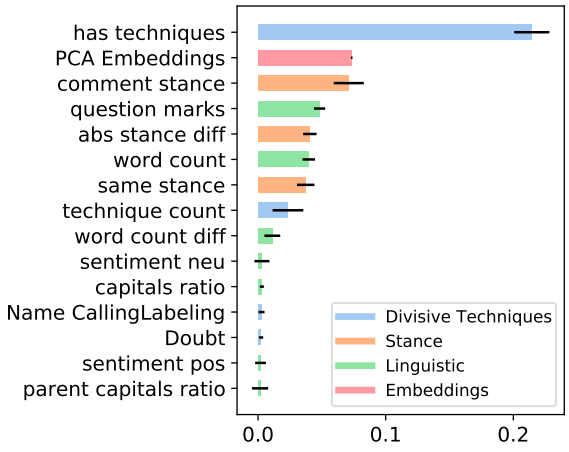


Figure 5: Feature importance analysis using permutation method.

jectives: reasoning models (like DeepSeek-R1) are optimized for analytical tasks requiring multiple evidence sources, while chat models (DeepSeek-V3) excel with minimal inputs but struggle with feature-rich representations, evidenced by performance degradation when adding features (from 55.11 to 49.77 macro-F1). Notwithstanding its limitations with multi-feature integration, DeepSeek-V3 ('chat model') surprisingly excelled in identifying agreement relationships, outperforming others in both destructive and constructive links detection, getting the best overall performance score (see Figure 8b in Appendix). This is likely due to its proficiency in interpreting base text and stance information, resembling their primary training objective – understanding conversational dynamics. Additionally, this advantage may also stem from an emphasis on contextual understanding over analytical depth, where these models prioritize comprehending nuances over handling complex, multi-faceted analyses, thereby facilitating their superiority in recognizing certain relationships, such as agreements.

6.1.1 Ablation Study

To identify the key predictors of divisiveness in online discussions, we conducted an ablation study using permutation importance. This technique measures feature importance by randomly shuffling each feature's values and calculating the resulting decrease in model performance, thus quantifying each feature's contribution to prediction accuracy independent of model architecture. We performed the analysis on our best-performing model—an XGBoost classifier using PCA-reduced embeddings (40 components) combined with linguistic, stance, and propaganda features. The permutation importance was calculated using 5 random permutations per feature on the test set, with macro F1 score as performance metric. Figure 5 presents the top features ranked by permutation importance, color-coded by category (blue for divisive techniques, orange for stance features, green for linguistic features, and pink for embeddings). Our analysis reveals a clear hierarchy in feature importance, with two features demonstrating substantially higher influence than others:

Divisive techniques: The binary indicator of whether a comment employs propaganda techniques (`has_techniques`) emerged as the strongest predictor of divisiveness (0.22 ± 0.01), suggesting that rhetorical manipulation strongly correlates with destructive discourse.

Comment stance: The ideological position expressed in a comment towards the topic (`comment_stance`) represents the second most influential feature (0.07 ± 0.01), indicating its importance in determining agreement/disagreement relationships between comments, which constitutes one of the two dimensions in our annotation schema.

Secondary predictors include `question_marks` (0.05), `abs_stance_diff` (0.05), and `word_count` (0.04), demonstrating the role of linguistic patterns and stance differences in predicting divisiveness. We grouped all the embedding components in just one indicator (`pca_Embeddings`), which appears among the most important features. This indicates that semantic content captured by contextualized representations contributes additional predictive power beyond explicit features.

The results provide evidence supporting the theoretical distinction between constructive and destructive communicative processes proposed by Samson and Nowak (2010), as well as our operationalization of the construct using divisive rhetoric and the specific stance expressed. This finding further validates our hybrid approach combining explicit rhetorical and stance features with semantic ones.

6.2 Chain Level

The chain-level analysis reveals intriguing methodological insights about how model performance transfers across analytical levels. As described in Section 5, we derived chain-level predictions by averaging comment-level scores from our best-performing models, mapping each chain to one of the categories defined in Table 1. While XGBoost

excels at the comment level, SVM unexpectedly performs better at the chain level. This counter-intuitive result likely stems from the interaction between class distribution and error patterns across analytical levels: destructive comments (57.48% of the dataset) appear in more balanced proportions when aggregated into chains, the models' error distributions affect chain-level metrics differently. SVM's marginal advantage in classifying destructive agreement comments becomes amplified when predictions are averaged into chain scores. This finding suggests that model selection should prioritize the specific analytical level of interest rather than assuming performance transfers across levels, as optimal classification at one level does not necessarily translate to optimal performance when those classifications are aggregated into higher-level constructs. A similar pattern emerges with LLMs, where OpenAI o3-mini shows improved performance at the chain level, suggesting how error distribution can impact model effectiveness across different analytical levels.

7 Discussion and Conclusions

This study addressed the challenge of automatically detecting and measuring constructive versus destructive communication patterns in online discussions. Starting from the theoretical framework proposed by Samson and Nowak (2010), we operationalized these constructs through a multilevel analytical approach examining both individual comments and conversation chains.

Given the abstract nature of these concepts, we extracted linguistic, stance and rhetorical features to characterize comments and highlight their communicative qualities. Our findings demonstrate the effectiveness of this theory-driven feature engineering approach. In fact, the ablation study revealed that divisive rhetorical techniques and stance information serve as the strongest predictors of destructive communication, substantially outperforming semantic embeddings alone.

This highlights an important methodological insight: when equipped with theoretically-grounded, specialized features, traditional machine learning approaches outperformed more complex models in domain-specific task. While LLMs excel at general language understanding, their performance is constrained when analyzing nuanced rhetorical and dialogical relationships that require explicit theoretically-grounded representation. The XGBoost model achieved 60.60% macro-F1 at the comment level, substantially outperforming the best LLM (DeepSeek at 55.11%). This advantage was even more pronounced at the chain level, where SVM reached 75.28% macro-F1 compared to OpenAI o3-mini's 64.69%. This gap underscores how domain-specific tasks requiring specialized theoretical knowledge may present unique challenges for general-purpose LLMs, which lack explicit representation of the theoretical structures provided by our feature engineering approach.

In conclusion, our study demonstrates the benefit of combining theoretical frameworks with computational methods for more nuanced approaches to controversy analysis. Establishing baselines for this task, and releasing our dataset and scripts, we aim to facilitate further exploration of how specific communicative choices contribute to either productive dialogue or increased antagonism across different platforms and domains, ultimately shaping conversational dynamics in online spaces.

Limitations

Our study has several limitations. LLMs used for rhetorical technique identification may underperform in domains different from their training data, struggling with detecting fallacious arguments "in the wild" (Ruiz-Dolz and Lawrence, 2023). Moreover, the qualitative analysis of rea-

Model	Base	B+Stance	B+Tech	B+S+T	B+Emb	B+S+T+E
DeepSeek (Chat)	41.95	44.70	44.45	49.72	-	-
DeepSeek (Reasoning)	38.53	43.80	47.17	49.89	-	-
GPT-mini	40.51	41.47	30.32	41.28	-	-
OpenAI o3-mini	**64.69**	59.38	62.99	59.57	-	-
Logistic Regression	24.66	39.43	62.26	64.61	36.41	71.72
SVM	26.94	35.42	62.80	66.67	40.70	**75.28***
XGBoost	26.18	46.83	64.67	69.47	38.68	70.18
Random Forest	25.00	37.86	70.15	68.10	39.06	69.72

*Best model overall for chain-level category prediction

Table 4: Chain-Level Divisiveness Detection Performance Comparison (F1 Macro Scores)

Chain Category	P	R	F1
Highly Destructive	0.83	0.83	0.83
Moderately Destructive	0.76	0.83	0.79
Slightly Dest./Neutral	0.67	0.71	0.69
Constructive	0.89	0.57	0.70
Macro Avg.	0.79	0.74	0.75

Model: SVM on All Features,
PCA 45; C: 1.0, gamma: 0.01, kernel: rbf

Table 5: Chain-Level Class-Specific Performance

soning models suggests that improved prompting strategies could potentially enhance LLM performance, indicating that our findings may reflect current implementation limitations rather than fundamental model constraints.

The dataset's skew toward destructive comments (57.48%), while reflective of "natural" YouTube discourse patterns, biases classification despite SMOTE implementation. Through more balanced datasets models equally sensitive to both constructive and destructive patterns could be developed. Additionally, moderate inter-annotator agreement ($Cohen's K = 0.37$) reflects the inherent subjectivity in evaluating discourse quality. Refined annotation protocols could improve gold standard roboustness in future work.

Our analysis focused exclusively on textual features, overlooking valuable structural information in conversation chains. Additionally, our chain-level predictions were derived by averaging comment-level scores. Incorporating graph-based features such as reply depth, branching and temporal patterns could enhance prediction performance, particularly for chain-level analysis (De Kock and Vlachos, 2021; Hessel and Lee, 2019).

The study's scope is confined to a single platform, language, and topic domain, limiting generalization. Cross-platform validation across diverse languages and topics is necessary for broader applicability.

Acknowledgments

This project has received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

Kelsey Allen, Giuseppe Carenini, and Raymond Ng. 2014. Detecting Disagreement in Conversations using Pseudo-Monologic Rhetorical Structure. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024a. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Davide Bassi, Michele Joshua Maggini, Renata Vieira, and Martín Pereira-Fariña. 2024b. A pipeline for the analysis of user interactions in youtube comments: A hybridization of llms and rule-based methods. In *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 146–153.

K. Chen, Z. He, R.-C. Chang, J. May, and K. Lerman. 2023. Anger Breeds Controversy: Analyzing Controversy and Emotions on Reddit. *Lecture Notes in Computer Science*, 14161:44–53.

M. Coletto, K. Garimella, A. Gionis, and C. Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3:22–31.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online. Association for Computational Linguistics.

K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. 2016. Quantifying controversy in social media. *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 33–42.

Randy Allen Harris, Chrysanne Di Marco, Sebastian Ruan, and Cliff O?Reilly. 2018. An annotation scheme for rhetorical figures. *Argument and Computation*, 9(2):155–175.

Jack Hessel and Lillian Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1648–1659.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Julia Jose and Rachel Greenstadt. 2024. Are large language models good at detecting propaganda? In *Proceedings of the ICWSM Workshops*.

Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3899–3906.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

John Lawrence, Mark Snaith, Barbara Konat, Katarzyna Budzynska, and Chris Reed. 2017. Debating Technology for Dialogical Argument: Sensemaking, Engagement, and Analytics. *ACM Transactions on Internet Technology*, 17(3):24:1–24:23.

Noortje Marres. 2015. Why map issues? on controversy analysis as a digital method. *Science, Technology, & Human Values*, 40(5):655–686.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Katarzyna Samson and Andrzej Nowak. 2010. Linguistic signs of destructive and constructive processes in conflict. In *IACM 23rd Annual Conference Paper*.

Jennifer Schumann and Steve Oswald. 2024. Pragmatic perspectives on disagreement. *Journal of Language Aggression and Conflict*, 12(1):1–16.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

A. Sriteja, P. Pandey, and V. Pudi. 2017. Controversy detection using reactions on social media. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017:884–889.

Robin R. Vallacher, Peter T. Coleman, Andrzej Nowak, Lan Bui-Wrzosinska, Larry Liebovitch, Katharina Kugler, and Andrea Bartoli. 2013. *Attracted to Conflict: Dynamic Foundations of Destructive Social Relations*, 1 edition. Peace Psychology Book Series. Springer Berlin, Heidelberg, Berlin.

Douglas Walton. 2008. *Informal logic: A pragmatic approach*. Cambridge University Press.

H. Wang, Y. Wang, X. Song, B. Zhou, X. Zhao, and F. Xie. 2023. Quantifying controversy from stance, sentiment, offensiveness and sarcasm: a fine-grained controversy intensity measurement framework on a Chinese dataset. *World Wide Web*, 26(5):3607–3632.

Joseph Zompetti. 2015. *Divisive discourse*. Cognella Academic Publishing, Illinois State University.

The gold-set for evaluating stance detection performance was created by two annotators who independently labeled 1,300 comments (guidelines in repository), achieving an inter-annotator agreement of Cohen's $\kappa = 0.61$. Disagreements were resolved through discussion, and GPT-4o's performance was subsequently evaluated on this gold-set; performance metrics are reported in Table 6.

Class	Precision	Recall	F1-score	Support
Against	0.833	0.743	0.785	502
Neutral/Other	0.602	0.730	0.660	400
Support	0.823	0.759	0.790	403
Macro	*0.752*	*0.744*	*0.745*	*1305*
Weight. Avg.	–	–	*0.748*	*1305*

Table 6: Performance metrics of the stance classification

For divisive rhetoric detection performance, one single expert annotator manually checked ChatGPT-4o-mini predictions on 2715 comments (see repository), following prompt definitions. Table 7 reports the performance metrics.

Technique	Prec.	Rec.	F1	Support
Overall Performance				
Micro Average	0.840	0.797	0.818	2175
Macro Average	0.791	0.659	0.696	-
Individual Techniques				
Appeal to Authority	0.652	0.577	0.612	26
Appeal to Fear/Prejudice	0.840	0.748	0.791	119
Bandwagon	0.667	0.200	0.308	10
Black-and-White Fallacy	0.828	0.485	0.611	99
Causal Oversimplification	0.676	0.881	0.765	227
Doubt	0.852	0.762	0.805	227
Exaggeration/Minimisation	0.862	0.880	0.871	241
Flag-Waving	0.882	0.833	0.857	108
Loaded Language	0.915	0.966	0.940	443
Name Calling/Labeling	0.869	0.896	0.883	415
Repetition	0.571	0.462	0.511	26
Slogans/Thought-terminating Cliché	0.821	0.222	0.350	149
Whataboutism/Straw Men	0.848	0.659	0.742	85

Table 7: Performance metrics of the divisive rhetorical techniques detection

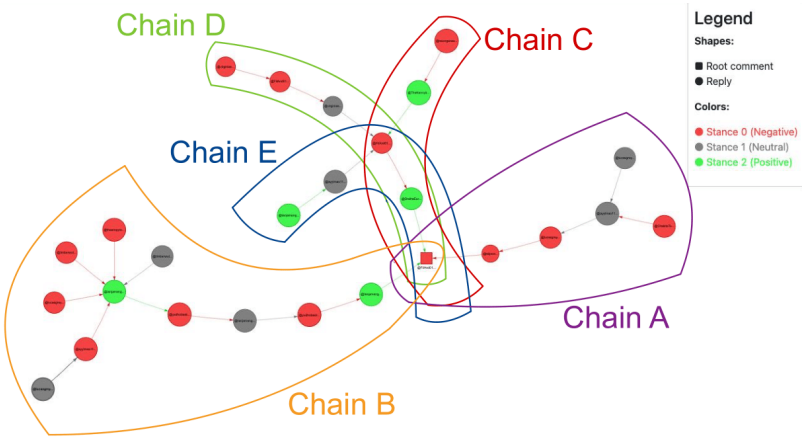
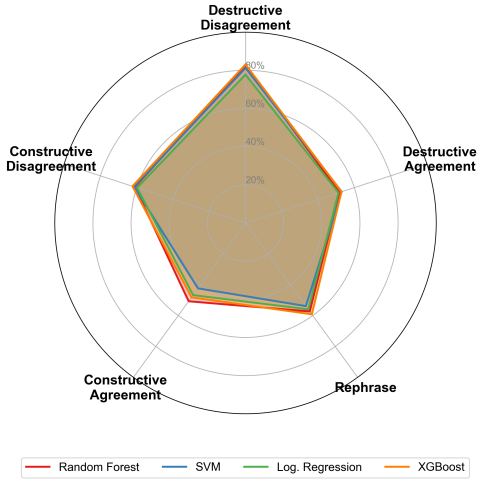
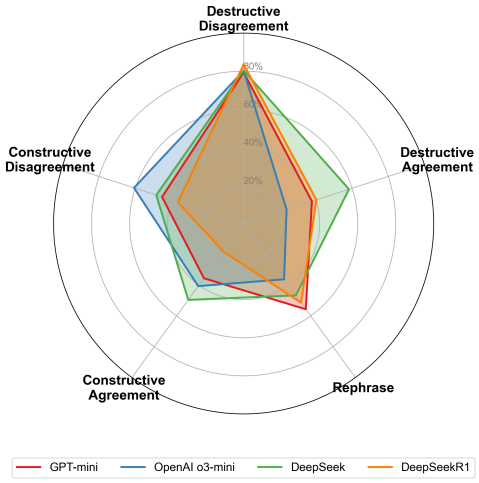


Figure 6: Clustering Example

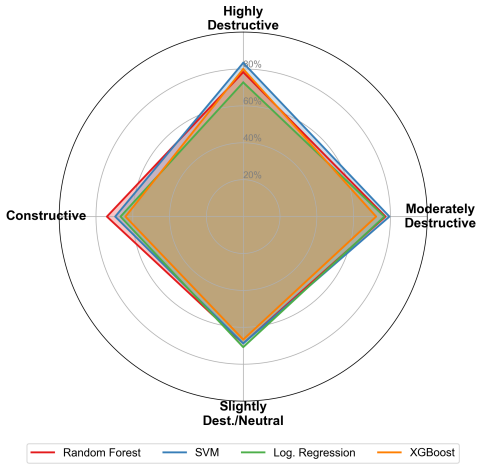


(a) Performance comparison of traditional machine learning models across all five categories.

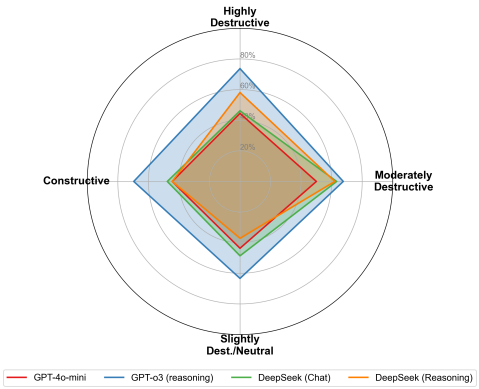


(b) Performance comparison of large language models across all five categories.

Figure 7: F1 score performance comparison of different model types on the five controversy classification categories.



(a) Performance comparison of traditional machine learning models across all chain categories.



(b) Performance comparison of large language models across all chain categories.

Figure 8: F1 score performance comparison of different model types on the four chain classification categories.