A Two-Axis Framework for Analyzing Ukrainian Dialogues

Artem Korotenko
Kyiv School of Economics
akorotenko@kse.org.ua

Roman Kyslyi
Kyiv School of Economics
rkyslyi@kse.org.ua

Abstract

Online discussions increasingly serve as a major venue for exchanging information and evaluating competing viewpoints. Yet most computational approaches to discourse quality focus on detecting harmful language or predicting engagement, providing limited insight into whether interactions actually improve collective understanding.

We introduce an initial two-dimensional framework for modeling dialogic constructiveness, distinguishing between substantive contribution (SC) and relational conduct (RC). Using expert-annotated Ukrainian-language discussions, we find preliminary evidence that collapsing rubric-level labels into these axes improves inter-annotator agreement, which is consistent with constructiveness being captured more reliably as a multidimensional judgment.

We further compare nominal, regression, and ordinal prediction approaches and find that explicitly modeling constructiveness as an ordinal task yields higher agreement with expert annotations under quadratic weighted kappa (QWK) on our gold test set. These results are consistent with dialogic constructiveness being more effectively modeled as an ordered interactional judgment than as a binary label or continuous score.

1 Introduction

Online discussions increasingly shape how people understand political, social, and scientific issues. Comment threads on news sites, forums, and messaging platforms often serve as a primary space for exchanging information and evaluating competing viewpoints. However, while some discussions help participants clarify arguments or reconsider positions, others produce confusion, repetition, or conflict without contributing to shared understanding.

Current computational tools can detect harmful language or predict engagement, but they provide limited insight into whether a discussion is actually productive. In practice, conversations that generate new insight and those that merely generate noise may appear equally civil or equally active. This makes it difficult to assess whether online interaction supports collective reasoning or simply amplifies disagreement.

The concept of *constructiveness* attempts to capture this positive dimension of discourse quality. In ordinary language, *constructive* means "serving a useful purpose" or "helping to improve something rather than damage it" (Oxford University Press, 2026). In interactional terms, a constructive comment is one that helps move a discussion forward instead of obstructing it. Kolhatkar and Taboada (2017) define constructive comments as follows:

> "Constructive comments intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. They are typically targeted to specific points and supported by appropriate evidence."

However, in most computational work, constructiveness is treated as a single binary label (Napoles et al., 2017; Kolhatkar et al., 2020; Nguyen et al., 2021). A comment is classified as either constructive or non-constructive. Empirically, such labels show only moderate inter-annotator agreement, suggesting that annotators compress multiple dimensions of discourse quality into one scalar judgment. Conceptually, this formulation conflates interpersonal conduct with argumentative substance.

This binary approach leaves a vast *"grey zone"* of interaction unaddressed: the cognitively demanding but relationally rough disagreements where participants may express strong opposing views without resorting to *ad hominem* attacks. To bridge this gap, we propose a shift from binary classification to a dual-axis architecture.

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 24–32

We define constructiveness as a multidimensional property composed of two independent components:

1. Relational Conduct (RC) — the interpersonal quality of interaction, including respect, engagement, and conflict handling (Gibb, 1961; Gottman, 1994)

2. Substantive Contribution (SC) — the epistemic value of the message, including reasoning quality, justification, informativeness, and coherence within the dialogue (Habermas, 1984; Grice, 1975; Steenbergen et al., 2003).

These dimensions are orthogonal. A contribution may be respectful yet add little substance, or it may present rigorous reasoning while escalating interpersonal tension. Collapsing both into a single label obscures this structure and limits analytical precision.

For this reason, we move from binary classification to a two-axis framework that disentangles Relational Conduct from Substantive Contribution, allowing a more granular analysis of conversational health.

We demonstrate the utility of this framework by applying it to a novel dataset of Ukrainian political dialogues sourced from **Telegram** and **Ukrayinska Pravda Forum**. By utilizing a 10-message context window, we capture pragmatic signals—such as sarcasm and relational repair—that are frequently lost in single-comment analysis.

We present this work as an exploratory study intended to motivate further investigation. Our contributions are as follows:

1. We introduce a two-axis theoretically grounded framework for conversational health with six specific subdimensions.

2. We show that collapsing rubric-level labels into these axes improves inter-annotator agreement, suggesting more stable latent dimensions of discourse quality.

3. We demonstrate that modeling constructiveness as an ordinal prediction task yields higher agreement with expert judgments than nominal or continuous formulations.

2 Related Work

The automated analysis of online discourse has produced several foundational resources, yet existing frameworks remains largely fragmented across single-label or outcome-based metrics.

1. **The Yahoo News Annotated Comments Corpus** (YNACC) attempted to identify "constructive" comments but found the category too broad for consistent human judgment, resulting in a Krippendorff's alpha of only 0.48–0.63 (Napoles et al., 2017)

2. **The Constructive Comments Corpus** (C3) (Kolhatkar et al., 2020) labeled comments in isolation, a method that often fails to detect pragmatic signals like sarcasm or "repair attempts" that only emerge across multiple conversational turns;

3. **The ChangeMyView** (CMV) dataset measures constructiveness through the lens of persuasion (Tan et al., 2016), focusing on the outcome (successful change of mind) rather than the deliberative process itself.

4. **The Stanford Politeness Corpus** utilizes computational markers (e.g., hedges, gratitude) to score etiquette (Danescu-Niculescu-Mizil et al., 2013), which can lead to the "Kind but Empty" trap where civil but substantively vacuous content is amplified

5. Recent work by Zhou et al. (2024) utilizes Large Language Models (LLMs) to extract dataset-independent linguistic features—such as dispute tactics and collaboration markers—to predict constructiveness outcomes . While this framework improves model interpretability, it highlights a persistent issue: the omission of a formal, *a priori* definition of constructiveness . By defining "constructiveness" purely through downstream proxies—such as the avoidance of mediation (escalation) or self-reported open-mindedness scores—these models risk learning robust prediction rules for specific datasets while failing to generalize a universal standard for conversational health. Without a top-down theoretical architecture, bottom-up feature extraction remains bound to the artifacts of the target variable it seeks to predict

3 Proposed Framework

The guiding hypothesis of this framework is that constructiveness may not be adequately captured

by a single score, because it appears to conflate two conceptually distinct dimensions. **Relational Conduct (RC)** measures the interpersonal quality of treatment among participants, rooted in Gottman's and Gibb's theories. **Substantive Contribution (SC)** measures the rational-deliberative quality of the content, grounded in the Deliberative Quality Index and deliberative systems theory.

3.1 Scoring Mechanics: The 5-Point Gradient

Every subdimension is evaluated on a centered ordinal scale from -2 to $+2$. A five-point scale was selected as a compromise between expressiveness and annotation reliability. Simpler three-level schemes cannot capture meaningful variation in conversational quality, while more granular scales introduce annotation noise due to small perceptual differences between adjacent categories. The five-point structure provides a clear neutral anchor (0) for factual but shallow contributions while allowing two degrees of positive and negative constructiveness that better match human perception of conversational intensity.

Score	Description
+2	Exemplary: perspective-taking, structured reasoning, proactive de-escalation
+1	Constructive: polite engagement and clear reasoning
0	Neutral: factual and professional but shallow
-1	Poor: dismissiveness, passive-aggressiveness, or topical drift
-2	Anti-constructive: insults, mockery, or escalation

The initial rubric contained eight subdimensions—four relational and four substantive—designed to capture a broad range of dialogic behaviors. In practice, however, annotation revealed substantial overlap between closely related criteria. Real-world comments often exhibited signals that were difficult to attribute to a single dimension, and annotators struggled to consistently separate concepts such as procedural fairness from reasoning quality or responsiveness from discussion progress.

Following several calibration rounds, we simplified the rubric by merging conceptually adjacent dimensions. The resulting framework preserves the two-axis structure but reduces each axis to three core subdimensions, yielding a six-dimensional scheme. This revision maintains theoretical coverage while improving annotation clarity and inter-annotator consistency.

3.2 Relational Conduct (RC): The Interpersonal Layer

Relational Conduct refers to how participants treat one another within an exchange. It captures whether interlocutors recognize each other as legitimate participants in dialogue and whether disagreement is handled in a way that sustains, rather than damages, the interaction.

This dimension is grounded in established communication research. Gibb's theory of defensive versus supportive communication distinguishes interactional climates that encourage openness from those that trigger defensiveness (Gibb, 1961). Gottman's work on conflict behavior identifies contempt and hostility as reliable indicators of relational breakdown (Gottman, 1994). Politeness theory further conceptualizes interaction in terms of managing face threats and acknowledging the social standing of the interlocutor (Brown and Levinson, 1987). Across these traditions, the central concern is not agreement but the preservation of interactional conditions necessary for continued dialogue.

Relational Conduct therefore evaluates whether a message maintains the possibility of discussion. It does not measure emotional neutrality or consensus. Strong disagreement may still exhibit high relational quality if expressed without personal attack or delegitimization. Conversely, even factually correct statements may score low if delivered in a way that undermines the interaction itself.

In our framework, Relational Conduct is decomposed into three components: tone and respect (RC1), engagement with the interlocutor (RC2), and conflict management (RC3). Together, these capture whether the relational layer of the conversation remains intact.

3.3 Substantive Contribution (SC): The Deliberative Layer

The Substantive axis assesses the intellectual capital and "logical payload" added to the conversation.

The framework is rooted in Jürgen Habermas's concept of an "ideal speech situation," where communication is undistorted by power or manipulation and participants are swayed only by the "unforced force of the better argument" (Habermas, 1984)

These Habermasian ethics is operationalized through the lens of the Discourse Quality Index (DQI) (Steenbergen et al., 2003), an empirical methodology originally designed to measure the

Axis	Subdimension	Definition
Relational Conduct (RC)	RC1. Respect / Tone	Focuses on emotional attitude toward others. Measures politeness, hostility, and warmth of language, without judging reasoning or topic relevance.
	RC2. Engagement	Captures social awareness and participation in dialogue. Evaluates whether the speaker actively engages with others and shows interest in continuing the exchange.
	RC3. Conflict Management	Reflects how disagreement is handled. Assesses whether the message escalates tension, stays neutral, or seeks understanding and compromise.
Substantive Contribution (SC)	SC1. Reasoning Quality	Assesses logical structure and fairness of argumentation. Focuses on how claims are supported by reasoning and whether evidence or opposing views are treated honestly.
	SC2. Informativeness	Measures content value. Evaluates how much new, relevant, or specific information the comment adds, separate from tone or reasoning.
	SC3. Dialog Continuity	Captures how well a comment fits into the ongoing dialogue. Evaluates whether it logically follows from what was said before and contributes to the development of the discussion instead of breaking or diverting its flow.

Table 1: The Two-Axis Framework Scoring Rubric

quality of parliamentary deliberation. The DQI focuses heavily on the "level of justification," providing a hierarchy that distinguishes between claims made with no support, inferior justification involving tenuous or fallacious links, and sophisticated justification where multiple reasons are examined in depth.

While the DQI offers a robust foundation, its parliamentary origin assumes a structured environment with a predefined behavioral "floor". In the "wild" digital sphere, this scale lacks the granularity to detect passive-aggressive condescension or the "repair attempts" essential for de-escalation. Moreover, the DQI's three-to-four level justification scale is too coarse to distinguish "polite" vacuity from "rough" but substantively deep contributions.

To resolve these limitations, the Substantive Contribution (SC) axis decomposes value into three granular sub-dimensions. SC1 (Reasoning Quality and Integrity) shifts focus from subjective agreement to objective deliberative rigor and epistemic honesty. SC2 (Informativeness) operationalizes Gricean maxims (Grice, 1975) by rewarding original insights that transform the thread while penalizing "empty" content. Finally, SC3 (Dialog Continuity) maintains systemic health by evaluating semantic coherence and penalizing irrelevant non-sequiturs that stall the deliberative flow (Mansbridge and Parkinson, 2012).

4 Dataset Collection and Annotation

4.1 Sourcing: Telegram and Ukrayinska Pravda Forum

We curated a novel dataset of Ukrainian political discourse [1] by scraping two distinct environments that represent different eras and modes of online interaction:

- **Telegram (TG):** We targeted ShrikeChat[2], one of the most prominent Ukrainian political discussion hubs.
- **Ukrayinska Pravda (UP) Forum[3]:** A legacy platform with over 20 years of history, capturing the long-term evolution of the Ukrainian digital public sphere.

Unlike previous datasets that group comments into loose threads, our methodology preserves the **direct-response chain**.

1. **Queue Extraction:** Linear sequences where each message is an explicit reply to the preceding one.
2. **9-Message Context Window:** For every target message, the nine preceding direct responses were preserved. Recent findings in

[1] https://huggingface.co/datasets/KSE-RESEARCH-Group/ukrainian-dialogs-constructiveness

[2] https://t.me/shrikechat

[3] https://forum.pravda.com.ua/index.php?board=2.0

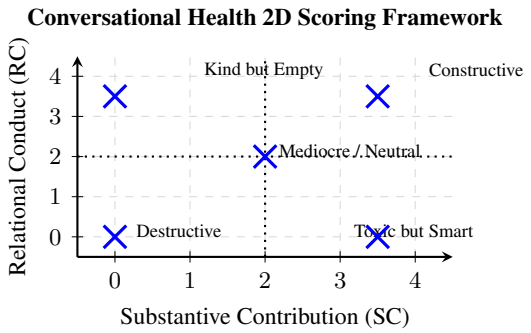

Figure 1: Two-axis conversational health scoring framework.

> NLP show the importance of context to grasp the nuances of conversation, specifically regarding how perceived toxicity can shift when the preceding dialogue is considered (Xenos et al., 2022)

This approach enables distinction between unprovoked aggression and retaliatory low-RC scores, as well as identification of substantive repair.

4.2 Labeling and Annotation

On the first stage, each dialogue was pre-labeled using the Gemini 2.5 Flash model (Google DeepMind, 2025). Manual annotation was subsequently performed using the Argilla framework (Vila-Suero and Aranda). Due to technical constraints, the theoretical -2 to $+2$ scale was mapped to a 0 to 4 integer scale.[4]

Before constructing the final evaluation sets, calibration rounds were conducted to align human perception with the framework. Items with directional disagreement were discussed until consensus or documented divergence was reached.

The Gold Set was sampled from the pre-labeled corpus using a stratified approach.

Sampling prioritized dimensional coverage, bucket balancing, and stratum diversity, including rare cases such as "Toxic but Smart" (low RC, high SC).

5 Data Analysis

To validate the reliability of our multi-dimensional annotation framework, we conducted an extensive Inter-Annotator Agreement (IAA) analysis. The dataset consists of 1,371 annotated items divided into two subsets. The gold subset was annotated independently by the same three calibrated expert annotators, providing a consistent expert reference set used for evaluation and reliability analysis. The main annotation set was labeled by two annotators per item, with a subset receiving a third annotation when additional verification was required. This design allows the main dataset to increase training coverage while maintaining a stable expert-validated subset for benchmarking and agreement analysis.

We assess reliability through three complementary perspectives: agreement at the level of individual rubric axes, agreement after collapsing axes into the two higher-level meta-dimensions (Relational Constructiveness and Substantive Constructiveness), and the empirical independence of the relational and substantive rubrics.

5.1 Inter-Annotator Agreement (IAA)

Table 2: Inter-Annotator Agreement (IAA) for the calibrated expert annotators ($N = 300$). Meta-dimensions collapse granular axes into a 2D framework, showing significantly higher conceptual alignment.

Dimension	QWK	α	Ex.%	Cl.%
RC1: Respect/Tone	0.650	0.647	68.0	96.8
RC2: Engagement	0.471	0.445	49.8	90.8
RC3: Conflict Mgmt	0.555	0.562	64.6	95.7
SC1: Reasoning	0.611	0.604	56.8	92.9
SC2: Informativeness	0.682	0.674	54.2	93.4
SC3: Dialog Continuity	0.521	0.512	58.8	93.0
Content (Y)	**0.761**	**0.762**	–	–
Relationship (X)	**0.666**	**0.663**	–	–

We report four agreement metrics per dimension. **Quadratic weighted Cohen's** κ (QWK) and **Krippendorff's** α are chance-corrected coefficients that penalize disagreements by squared distance on the ordinal scale; both range from -1 to 1, with higher values indicating stronger chance-corrected agreement. **Ex.%** denotes exact agreement and **Cl.%** close agreement (within ± 1).

5.2 2D Quality Framework and Orthogonality

One observation in our analysis is the transition from 6 granular axes to a 2D Quality Map. Collapsing the axes into meta-dimensions—Relationship ($X = \sum RC_{1-3}$) and Content ($Y = \sum SC_{1-3}$)—yields higher signal-to-noise ratios. The Content Meta score reached $\alpha = 0.801$, a level of agreement consistent with—though not by itself establishing—a more stable two-dimensional latent structure. The average Euclidean distance between expert points in

[4] A score of 2 represents the Neutral (0) anchor.

this 12×12 space is 1.66 units, indicating close agreement.

Furthermore, Pearson correlation analysis supports the orthogonality of these dimensions. As illustrated in the correlation matrix (Table 3), while internal clusters show moderate correlation ($r \approx 0.5$), cross-cluster correlation is significantly lower. Specifically, the correlation between *Respect/Tone* and *Informativeness* is weak ($r = 0.14$), justifying the treatment of model "politeness" and "accuracy" as independent variables.

Table 3: Pearson Correlation Matrix between annotation axes ($N = 300$). Shading indicates correlation strength: Darker blue denotes $r \geq 0.5$; lighter shades denote $r < 0.3$.

	RC1	RC2	RC3	SC1	SC2	SC3
RC1	1.00	0.58	0.68	0.23	0.16	0.24
RC2	0.58	1.00	0.60	0.29	0.14	0.38
RC3	0.68	0.60	1.00	0.28	0.24	0.29
SC1	0.23	0.29	0.28	1.00	0.51	0.49
SC2	0.16	0.14	0.24	0.51	1.00	0.60
SC3	0.24	0.38	0.29	0.49	0.60	1.00

If constructiveness were a single scalar construct, relational and substantive indicators would collapse into one highly correlated cluster. Instead, we observe two moderately coherent but separable clusters, supporting the interpretation of constructiveness as a two-dimensional state.

6 Baseline Modeling

To examine whether the proposed constructiveness framework can be predicted automatically, we evaluate several baselines under different assumptions about label structure. Each subdimension (RC1–RC3, SC1–SC3) is modeled independently on a five-point ordinal scale (0–4) derived from the original -2 to $+2$ rubric. Performance is measured with Quadratic Weighted Kappa (QWK), which accounts for ordinal disagreement by penalizing larger errors heavily than near misses.

We use a fixed split protocol. The gold set (N=300) is divided into train and test (210/90). The gold training portion is merged with the main dataset, from which 15% is used for validation and 85% for training. Final evaluation is conducted exclusively on the gold test set (N=90).

All neural baselines share the same multilingual encoder, XLM-RoBERTa (Conneau et al., 2020), fine-tuned for constructiveness prediction. The nominal, regression, and ordinal models therefore differ only in output formulation and decoding strategy, making their comparison directly controlled.

Because QWK is sensitive to prediction variance, trivial strategies can achieve moderate proximity to the gold labels while still yielding near-zero agreement. We therefore calibrate all non-trivial baselines on the gold validation split before final evaluation.

The first baseline is a distribution-aware non-text model that samples labels from the empirical score distribution of the gold training subset for each axis. This captures dataset priors without using textual information.

We then train a nominal softmax classifier with cross-entropy loss, treating each axis as a five-class classification problem and ignoring the ordered structure of the scale. At inference time, class probabilities are converted to ordinal predictions using expectation decoding followed by rounding.

Next, we consider a regression baseline trained with mean squared error (MSE) to predict continuous constructiveness scores. Per-axis decision thresholds are tuned on the validation split to maximize QWK before testing. This allows the model to partially adapt to ordinal structure, although continuous predictions remain sensitive to boundary effects.

Finally, we evaluate an ordinal classifier based on CORAL (COnsistent RAnk Logits) (Niu et al., 2016). CORAL decomposes prediction into a sequence of threshold exceedance decisions and enforces monotonic consistency across class boundaries. As with regression, thresholds are calibrated on the validation split. By modeling ordered decision boundaries directly, CORAL is better suited to the ranked nature of the constructiveness scale.

6.1 Baseline Results

Table 4 summarizes the predictive performance of the evaluated baseline approaches on the held-out test split. In addition to Quadratic Weighted Kappa (QWK), we report the proportion of predictions within one ordinal step of the expert annotation (Near$\pm$1), which reflects coarse agreement with annotator judgments under the five-point constructiveness scale.

The distribution-aware baseline, which samples predictions from the empirical training distribution without using textual information, achieves relatively high Near$\pm$1 agreement despite near-zero QWK. This reflects the strong concentration of annotations within a narrow central range of the

Table 4: Baseline performance across six constructiveness dimensions. All models share an identical XLM-RoBERTa encoder and differ only in output formulation and decoding strategy.

Model	Macro QWK	Near±1	RC1	RC2	RC3	SC1	SC2	SC3
Distribution (no text)	−0.016	81.11%	0.019	0.018	-0.072	0.127	0.013	-0.116
Softmax (CE)	0.006	92.22%	0.008	0.000	0.030	0.000	0.000	0.000
Regression (MSE)	0.074	76.83%	0.122	0.059	-0.006	0.052	0.123	0.091
Ordinal (CORAL)	**0.3891**	**86.67%**	0.309	0.435	0.241	0.452	0.576	0.328

ordinal scale, where coarse proximity can be obtained without meaningful alignment with expert judgments.

The nominal softmax classifier substantially improves numerical proximity to the gold annotations, as reflected by lower MAE and higher Near±1 agreement. However, this formulation yields only marginal gains in QWK, indicating that while the model captures overall constructiveness polarity, it fails to reliably place predictions across the ordinal boundaries used by annotators.

The regression baseline partially recovers latent ordering information, resulting in modest improvements in ordinal agreement. Nevertheless, without calibrated decision thresholds, continuous predictions frequently cross class boundaries in ways that degrade alignment with the discrete annotation scale.

In contrast, the ordinal CORAL formulation achieves higher QWK across all dimensions while maintaining similar Near±1 agreement. This is consistent with the improvement being driven by explicit modeling of ordered structure rather than by changes in text representation. The ordinal model appears to better approximate the decision boundaries reflected in expert annotations within the proposed two-axis framework

7 Conclusion

We introduced a two-dimensional framework for modeling dialogic constructiveness in online discussions, separating substantive contribution (SC) from relational conduct (RC). Unlike prior approaches that treat constructiveness as a binary property of individual messages, the proposed formulation captures both the informational and interactional contributions of a message to the progression of a discussion.

Empirical analysis of expert annotations suggests that constructiveness judgments behave as multidimensional and ordered in our data. Collapsing the six rubric dimensions into the two proposed axes increases inter-annotator agreement, which is consistent with—though not conclusive of—RC and SC capturing more stable aspects of expert judgment than the individual annotation criteria.

Our modeling experiments further show that approaches ignoring ordinal structure struggle to align with expert decisions under agreement-based metrics. Nominal classifiers achieve high proximity to the gold labels but produce weak ordinal agreement, while regression models only partially recover the ordering of constructiveness scores. In contrast, explicitly modeling constructiveness as an ordinal prediction task yields higher alignment with expert judgments on our gold test set.

Overall, the results support modeling dialogic constructiveness as a bounded ordinal judgment expressed along multiple interactional dimensions, though further validation is needed. Future work should explore architectures that better exploit ordinal structure and extend the framework to additional languages, platforms, and forms of online discussion.

Ethics

The dataset was constructed from publicly available Telegram channels via the official Telegram API and from publicly accessible discussion threads on the Ukrainska Pravda Forum. No private groups, direct messages, or restricted-access content were accessed. Data were processed for research purposes under legitimate interest and scientific research provisions. Identifiers were pseudonymized at the time of data collection, and only content necessary for the analytical objectives was retained. Messages containing information that could reasonably enable re-identification of individuals were removed from the dataset prior to analysis.

Annotation and labeling were conducted by student researchers who receive educational grants that include a requirement to contribute a defined number of working hours to university research activities. Their participation in the annotation

process formed part of these structured research duties. Annotators were granted access exclusively to anonymized data and were instructed on confidentiality and data protection requirements prior to participation.

Limitations

Several limitations of this study should be noted.

First, the framework and models were developed using Ukrainian-language online discussions, primarily from political and civic contexts. Norms of disagreement and cooperation may vary across languages and communities, meaning that constructiveness thresholds learned in this setting may not directly transfer to other discourse environments.

Second, the current approach assumes an explicit reply-based conversational structure, evaluating each message in relation to its immediate dialogic context. In less structured discussions involving indirect or topic-level responses, constructiveness may depend on broader interaction patterns not captured by pairwise message adjacency.

Third, inter-annotator agreement remains moderate, particularly along relational dimensions. This reflects the subjective nature of tone and intent interpretation, and suggests that some learned decision boundaries may be influenced by annotation uncertainty.

Fourth, items were pre-labeled with Gemini 2.5 Flash before human review. Despite calibration rounds and independent expert annotation, we cannot rule out anchoring effects that may inflate observed agreement, particularly along the collapsed two-axis dimensions. Comparison with fully blind annotation is needed to quantify this.

Finally, the expert-annotated gold subset is relatively limited in size (N = 300, with N = 90 held out for test). Triple expert annotation across six ordinal dimensions is labor-intensive, which limited the feasible scale within our resources. While the larger main split (N = 1,071) supports broader training coverage, statistical power for the QWK comparisons reported on the gold test set remains limited; differences between baselines should be interpreted accordingly.

References

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *ACL Anthology*.

Jack R. Gibb. 1961. Defensive communication. *Journal of Communication*, 11(3):141–148.

Google DeepMind. 2025. Gemini 2.5 flash. Accessed: 2026-02-10.

John M. Gottman. 1994. *Why Marriages Succeed or Fail*. Simon & Schuster.

Herbert P. Grice. 1975. Logic and conversation. In *Speech Acts*, pages 41–58. Brill.

Jürgen Habermas. 1984. *The Theory of Communicative Action*. Beacon Press.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Varada Kolhatkar, Nanyun Wu, Luca Cavasso, Emmanuel Francis, and Maite Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Jane Mansbridge and John Parkinson. 2012. *Deliberative Systems: Deliberative Democracy at the Large Scale*. Cambridge University Press.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosendahl, and Madira Thampiratnam. 2017. Finding good conversations online: The yahoo news annotated comments corpus. *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. *Constructive and Toxic Speech Detection for Open-Domain Social Media Comments in Vietnamese*, page 572–583. Springer International Publishing.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928.

Oxford University Press. 2026. Constructive. https://www.oxfordlearnersdictionaries.com/definition/english/constructive. Accessed: 2026-02-10.

Marco Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in Good-faith online discussions. *arXiv preprint arXiv:1602.01103*.

Daniel Vila-Suero and Francisco Aranda. Argilla – open-source framework for data-centric NLP.

Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Léo Laugier. 2022. Toxicity detection sensitive to conversational context. *First Monday*, 27(5).

Lexin Zhou, Youmna Farag, and Andreas Vlachos. 2024. An LLM feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5389–5409. Association for Computational Linguistics.

A Annotation Examples from the Gold Set

A.1 Toxic but Smart

Original

"відомо про співвідношення переданих/відпущених хамасівців? чи просто чергове узагальнення" (vidomo pro spivvidnoshennia peredanykh/vidpushchenykh hamasivtsiv? chy prosto cherhove uzahalnennia)

Translation

Is there known data on the ratio of Hamas prisoners exchanged or released, or is this just another generalization?

Scores RC1 = -1, RC2 = 0, RC3 = -1; SC1 = 1, SC2 = 1, SC3 = 1

A.2 Kind but Empty

Original

"Дякую, а то я вже почала писати Краще не скажеш." (Diakuiu, a to ya vzhe pochala pysaty. Krashche ne skazhesh.)

Translation

Thanks, I had already started writing it myself. Couldn't have said it better.

Scores RC1 = 2, RC2 = 2, RC3 = 0; SC1 = 0, SC2 = -1, SC3 = 0

A.3 Constructive

Original

"Я це розумію, але я не розумію, чому не можна навести порядок в генштабі." (Ya tse rozumiiu, ale ya ne rozumiiu, chomu ne mozhna navesty poriadok v henshtabi.)

Translation

I understand that, but I do not understand why it is not possible to bring order to the General Staff.

Scores RC1 = 1, RC2 = 1, RC3 = 1; SC1 = 1, SC2 = 1, SC3 = 1

A.4 De-escalation

Previous message

"Та це ж очевидно: нормальні громадяни не голосують за Поршенка."

Response

"хто такі ці нормальні громадяни і чому невибір Поршенка зробив по вашому 73 % українців — ненормальними, та ще й 'прихильники певної політсили'? може досить бавитись в гівнокидання? питання дійсно серйозні." (khto taki tsi normalni hromadiany i chomu nevybir Porshenka zrobyv po vashomu 73% ukraintsiv nenormalnymy? mozhe dosyt bavytys v hivnokydannia? pytannia diisno seriozni.)

Translation

Who exactly are these "normal citizens," and why does not voting for Poroshenko make 73% of Ukrainians "abnormal" in your view? Maybe it is time to stop the mud-slinging; these are serious questions.

Scores RC1 = -1, RC2 = 1, RC3 = 2; SC1 = 1, SC2 = 1, SC3 = 1

A.5 Critical but Substantive

Original

"Тобто статистика показує, що Юкі був абсолютно неправий." (Tobto statystyka pokazue, shcho Yuki buv absoliutno nepravyi.)

Translation

So the statistics show that Yuki was completely wrong.

Scores RC1 = 0, RC2 = 0, RC3 = 0; SC1 = 1, SC2 = -1, SC3 = 1

A.6 Destructive

Original

"И ты такая же." (I ty takaya zhe.)

Translation

And you are the same.

Scores RC1 = -2, RC2 = -2, RC3 = -2; SC1 = -1, SC2 = -2, SC3 = -2

A.7 Neutral Inquiry

Original

"Це було до чи після початку повномасштабної війни?" (Tse bulo do chy pislia pochatku povnomasshtabnoi viiny?)

Translation

Was this before or after the start of the full-scale war?

Scores RC1 = 0, RC2 = 0, RC3 = 0; SC1 = 0, SC2 = 0, SC3 = 0