Enhanced Evaluative Language Annotation through Refined Theoretical Framework and Workflow

Jiamei Zeng[1*], Haitao Wang[2*], Harry Bunt[3], Xinyu Cao[2], Sylviane Cardey[4], Min Dong[5], Tianyong Hao[6], Yangli Jia[7], Kiyong Lee[8], Shengqing Liao[9], James Pustejovsky[10], François Claude Rey[4], Laurent Romary[11], Jianfang Zong[2], and Alex C. Fang[1**]

[1]City University of Hong Kong, PR China (jiamezeng3-c@my.cityu.edu.hk, acfang@cityu.edu.hk)
[2]China National Institute for Standardization, PR China ({wanght,caoxy,zongjf}@cnis.edu.cn)
[3]University of Tilburg, The Netherlands (Harry.Bunt@tilburguniversity.edu)
[4]University of Franche-Comté, France (sylviane.cardey@univ-fcomte.fr, francois_claude.rey@edu.univ-fcomte.fr)
[5]Beihang University, PR China (mdong@buaa.edu.cn)
[6]South China Normal University, PR China (haoty@m.scnu.edu.cn)
[7]Liaocheng University, PR China (jiayangli@lcu.edu.cn)
[8]Korea University, Korea (ikiyong@gmail.com)
[9]Fudan University, PR China (sqliao@fudan.edu.cn)
[10]Brandeis University, USA (jamesp@cs.brandeis.edu)
[11]National Institute for Research in Digital Science and Technology, France (laurent.romary@inria.fr)

Abstract

As precursor work in preparation for an international standard *ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16: Evaluative language*, we aim to test and enhance the reliability of the annotation of subjective evaluation based on Appraisal Theory. We describe a comprehensive three-phase workflow tested on COVID-19 media reports to achieve reliable agreement through progressive training and quality control. Our methodology addresses some of the key challenges through the refinement of targeted guideline refinements and the development of interactive clarification tools, alongside a custom platform that enables the pre-classification of six evaluative categories, systematic annotation review, and organized documentation. We report empirical results that demonstrate substantial improvements from the initial moderate agreement to a strong final consensus. Our research offers both theoretical refinements addressing persistent classification challenges in evaluation and practical solutions for the implementation of the annotation workflow, proposing a replicable methodology for the achievement of reliable annotation consistency in the annotation of evaluative language.

1 Introduction

Annotating evaluative language has become increasingly important in the present era of artificial intelligence, particularly given the need for high-quality language resources for training large language models and understanding human emotions and personal stance. In response to this need, we are working on an international standard (*ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16: Evaluative language*) for the annotation of evaluative language, to be adopted by the International Organization for Standardization. To ensure a sound practical application of the standard, we apply Appraisal Theory (AT) as a foundational framework of analysis, fully described in Martin and White (2005). While AT has proven influential across diverse research contexts, its practical implementation through manual annotation reveals significant methodological challenges. Research addressing annotation methodology remains scarce, with studies rarely reporting inter-coder agreement measures or addressing reliability issues (Fuoli, 2018). Similarly, the development of annotation workflow has received limited attention. While Fuoli (2018) proposed a stepwise approach to Appraisal annotation, the methodological guidance remains insufficient for addressing the complex practical challenges encountered in the annotation of evaluative language, which is fundamentally subjective. The limited focus on annotation methodologies has created a significant gap between theoretical significance and operational reliability, affecting the advancement of scientific interpretative approaches to the understanding of expressions of human stance and attitude.

We aim to address these methodological gaps through the development of annotation workflows, refinement of guideline procedures, and tool-facilitated quality control mechanisms. We describe a comprehensive approach that combines refined theoretical grounding with practical solutions for the achievement of reliable inter-annotator agreement based on the annotation of a corpus of authentic texts sampled from media reports. In what follows, we describe a methodological framework that encompasses a multi-stage training protocol, problem identification through pilot annotations, and targeted intervention strategies that reduce annotator disagreements.

[*] Equal contribution
[**] Corresponding author

Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21), pages 76–84

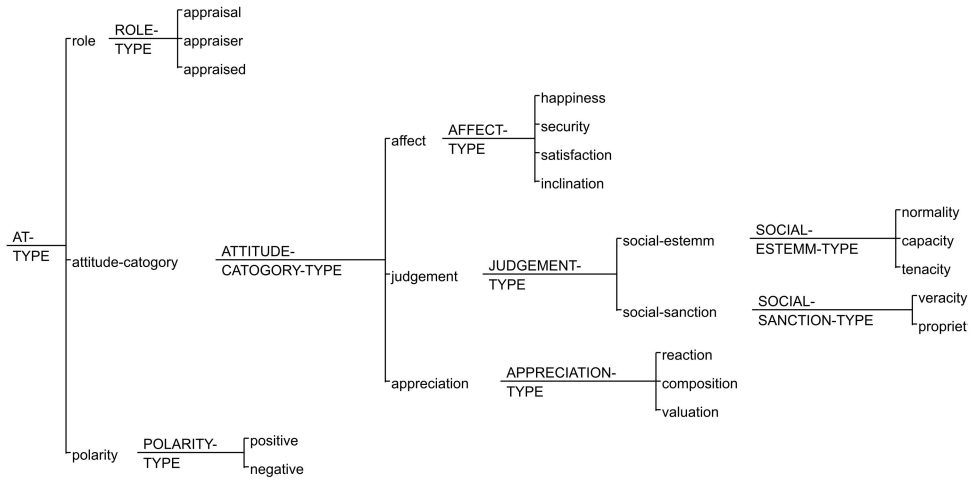

Figure 1: Annotation framework for Appraisal theory

2 Annotation Framework and Methodology

2.1 Annotation Framework

Our annotation framework is fundamentally grounded in AT fully described and published in Martin and White (2005) and specifically focuses on the Attitude system. In our approach, we explicitly define three core roles for each identifiable (markable) textual segment expressing an evaluation, namely, Appraiser (the entity making the evaluation), Appraised (the target being evaluated), and Appraisal (the appraisal element itself). The annotation process followed a hierarchical framework as shown in Figure 1.

2.2 Core Annotation Principles

Our annotation methodology is governed by two fundamental principles: (1) Minimality Principle, according to which annotators mark only the core evaluative lexical items and their essential modifiers, avoiding the unnecessary expansion of the textual segment, and (2) Completeness Principle, according to which the annotation segment must preserve the semantic integrity of evaluative expressions. Critical elements that influence evaluative meaning, particularly negation markers, must be included within the segment to maintain accurate interpretations and polarity determination. Consider

[1] *Research from earlier in the pandemic* ***does not yield definitive clues***. (USN0122)

In [1], while "definitive" represents the core evaluative term, the negation "does not" fundamentally alters the evaluative stance and must be included in the annotation span as "does not yield definitive clues" to preserve semantic completeness.

These principles establish baseline standards rather than rigid constraints. Recognizing the inherent complexity of the segment boundaries of evaluative language, we adopt a flexible approach to boundary matching for inter-annotator agreement assessment. Following Wiebe et al. (2005) and Read and Carroll (2012), overly strict boundary matching can be counter-productive. Therefore, when segments marked by different annotators overlap and demonstrate complete agreement on Main Category, Subcategory, and Polarity, we consider these as valid matches, specifically categorized as "Match with Overlap". This methodology maintains analytical rigour in categorical agreement while accommodating the subjective nature of evaluative language boundaries.

2.3 Corpus and Annotators

The texts used in our study are drawn from a corpus of COVID-19 news reports. The corpus comprises a total of 144 news articles related to COVID-19 balanced across four media outlets – China Daily, South China Morning Post, The Guardian, and The New York Times – with 36 articles from each. The articles were sampled from Factiva, covering the period from January 2020 to December 2022, with one article selected per month to ensure temporal balance. The descriptive statistics are presented in Table 1.

Articles were selected from this corpus to test the annotation of evaluative language, with each round

using articles from the same time periods across the four outlets, ensuring temporal and cross-media representation while maintaining manageable annotation loads.

	CD	SCMP	TG	NYT	Overall
Articles	36	36	36	36	144
Tokens	20536	23863	31432	42704	118535
Types	3710	4327	5062	6045	10701
TTR (%)	18.07	18.13	16.10	14.16	9.03
Mean size	570	663	873	1186	823

Table 1: Corpus composition and descriptive statistics across four media outlets

Two annotators were selected based on three criteria: (1) proficiency in English comprehension, (2) enrolment in a graduate-level programme in linguistics or English studies, and (3) foundational linguistic knowledge, including familiarity with Systemic Functional Linguistics (SFL). We intentionally selected annotators without prior independent experience in the annotation of evaluative language. The selection of novice annotators allows us to assess whether our framework can achieve satisfactory inter-annotator agreement through systematic training protocols, rather than relying on prior experience.

2.4 Annotation Software Tool

Many software tools facilitate basic annotation tasks but generally lack sophisticated comparison capabilities beyond simple agreement-disagreement identification. This limitation particularly affects the crucial intermediate review process, where detailed comparative analysis is essential for quality control and guideline refinement by providing organized documentation of annotator classifications and problematic and classic cases. Given these limitations, we adopt a hybrid approach combining existing and custom-developed tools. The initial annotation was conducted using the UAM corpus tool (O'Donnell, 2018). A web-based review tool has been designed and implemented that supports simultaneous upload of processed annotation results and enables detailed comparison of all annotation units within their original textual contexts, across the pair of annotators.

The tool also categorizes comparison results into six distinct types to provide fine-grained characterization of annotation consistency:

1. Annotator 1 Only: Annotations identified solely by Annotator 1
2. Annotator 2 Only: Annotations identified solely by Annotator 2
3. Match: Complete agreement in annotation content
4. Match with Overlap: Overlapping appraisal elements with identical classifications
5. Conflict: Identical text spans with different classification or polarity
6. Conflict with Overlap: Overlapping appraisal elements with different classifications or polarity

The interface enables comparative viewing of both annotators' work through individual or category-based review, with functions to flag typical or problematic cases and add comments. It identifies specific points of disagreement, supports targeted feedback, and offers traceable evidence for iterative guideline refinement. The tool also exports processed data for statistical analysis, significantly reducing manual comparison workload with minimal need for post-processing.

2.5 Statistical Analysis Methods

Our statistical approach is based on Read and Carroll (2012)'s evaluation methodology with a key modification. Read and Carroll (2012) note that "the 'number correct' (COR) will differ for each annotator in the pair under evaluation" due to multiple matching scenarios in text span annotation. We address this challenge through a different approach by implementing an explicit six-category classification system as a pre-processing step: Match, Match with Overlap, Conflict, Conflict with Overlap, Annotator 1 Only, and Annotator 2 Only, which represents a key functionality of our software tool, as detailed in the previous section. This approach simplifies statistical calculations by providing structured input for subsequent metrics while offering annotators and reviewers clear insight into agreement and disagreement patterns.

For all subsequent agreement calculations, including text anchor agreement, appraisal type agreement, and chance-corrected measures such as Cohen (1960)'s Kappa, we treat both Match and Match with Overlap categories as agreement instances. This approach recognizes that boundary variations do not necessarily indicate substantial disagreement in evaluation identification or classification, consistent with our flexible annotation principles outlined in Section 2.2.

3 Annotation Workflow Development

3.1 Training and Calibration Workflow

Our procedure for annotation training is designed as a systematic workflow to achieve reliable inter-annotator agreement through progressive skill development and quality control. The workflow encompasses three distinct phases, each serving specific methodological purposes in building up annotation competency.

The Foundation and Initial Practice phase establishes both theoretical knowledge and basic operational skills. Annotators begin with comprehensive study of the fundamentals of AT, followed by structured tutorials that bridge theoretical concepts and practical application. The initial practice involves supervised annotation of four articles with iterative feedback, ensuring that the annotators have developed proper software proficiency and terminological accuracy before proceeding to independent work in subsequent stages.

The Pilot Study and Problem Identification phase serves as both a competency test and a diagnostic tool. Eight articles were annotated independently by both annotators, generating comprehensive data for multi-dimensional statistical analysis. This phase is designed to reveal recurrent inconsistencies and conceptual confusions that require targeted intervention. The pilot study functions as a critical checkpoint to reveal specific areas where annotation guidelines need refinement and where annotators require additional training.

The Iterative Training and Quality Assurance phase addresses identified problems through targeted training cycles that alternate between collaborative learning and independent practice. Joint annotation sessions provide real-time guidance on problematic cases while independent practice allows for skill consolidation and performance monitoring. This iterative approach continues until the final assessment demonstrates that annotators can consistently achieve the targeted reliability threshold of 0.8 inter-annotator agreement.

3.2 Pilot Annotation and Performance Analysis

3.2.1 Initial Agreement Analysis

The pilot study produced 810 total annotation instances across both annotators, with substantial variations in annotated segments. Annotator 1 identified 520 segments of evaluation while Annotator 2 identified 631, indicating different thresholds for marking up evaluative language. Only 341 instances (42.1%) were commonly annotated, suggesting significant differences between the two. The distribution in Figure 2 reveals the extent of disagreement across the six-category system: Match (113 instances), Match with Overlap (21), Conflict (160), Conflict with Overlap (47), Annotator 1 Only (179), and Annotator 2 Only (290). The substantial proportions of unique annotations observed here highlight the scope of challenges in the annotation or, rather, the subjective interpretation of evaluative language.

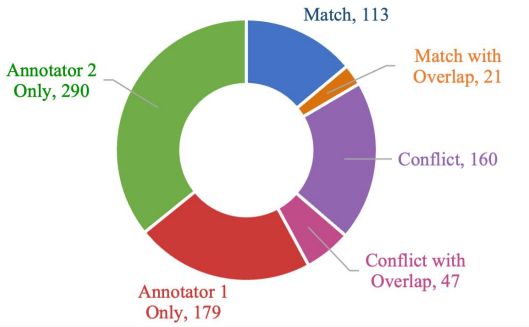

Figure 2: Distribution of annotation pairs (pilot study)

An agreement analysis using our extended classification framework revealed moderate consistency levels. The AGR values show asymmetric agreement patterns: AGR(1‖ 2) = 65.6% and AGR(2‖ 1) = 54.0%. These values are lower than comparable studies. For instance, Read and Carroll (2012) reported 70.6% and 68.6% respectively. This reflects significant disagreement between the two annotators regarding the identification of evaluative segments, reinforcing our initial suspicion that the interpretation of evaluative language is subjective and hence fundamentally controversial, a primary concern that triggered the present study in the first place.

	F_1	REC	PRE	ERR	UND	OVG
1 w.r.t. 2	0.233	0.212	0.258	0.835	0.460	0.344
2 w.r.t. 1	0.233	0.258	0.212	0.835	0.344	0.460
Mean	**0.233**	**0.235**	**0.235**	**0.835**	**0.402**	**0.402**

Table 2: MUC-7 test scores applied to all annotation instances (pilot study)

MUC-7 metrics (Chinchor, 1998) were applied to provide the detailed assessment in Table 2. Results indicate an overall F_1 score of 0.233 and a high error rate of 83.5%, revealing substantial room for improvement.

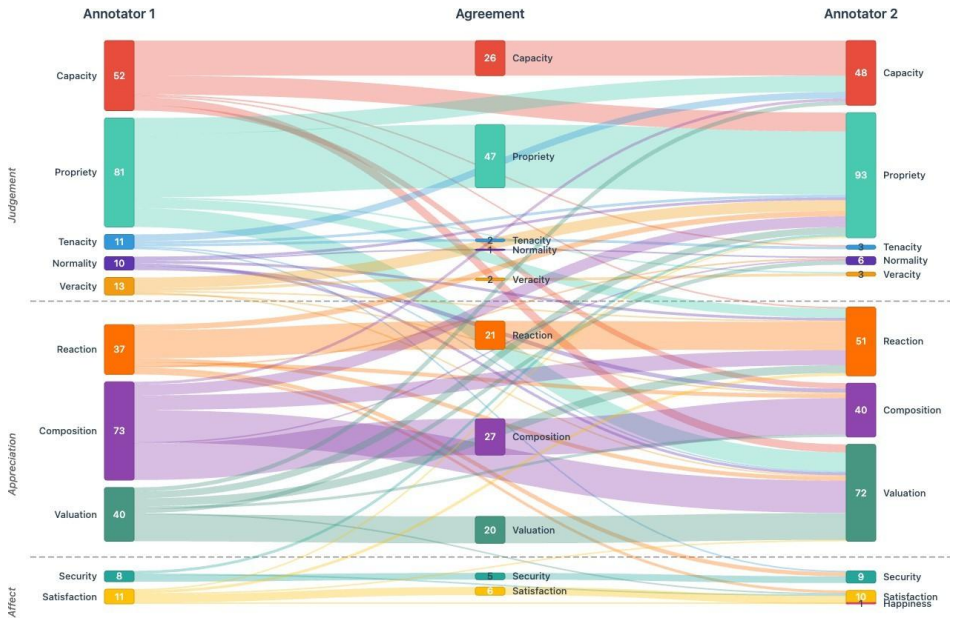

Figure 3: Annotation framework for Appraisal theory

3.2.2 Hierarchical Performance Analysis

To assess annotator agreement at different levels of complexity, we conducted a hierarchical analysis examining four progressive levels of classification requirements:

Level 0: Role identification (appraisal[1])
Level 1: Role + Main Category (Affect, Judgement, Appreciation)
Level 2: Role + Main Category + Subcategory
Level 3: Complete classification (Role + Main Category + Subcategory + Polarity)

Level	Overall	Affect	Judgement	Appreciation
1	0.551	–	–	–
2	0.360	0.478	0.429	0.506
3	0.354	0.400	0.380	0.420

Table 3: Cohen's Kappa values at different levels (pilot study)

Coohen's Kappa values, as presented in Table 3, show strong agreement at the main category level (κ=0.551) but moderate agreement for subcategories (κ=0.360) and complete classification (κ=0.354). Category-specific analysis reveals that Judgement subcategories were most problematic (κ=0.429), suggesting the need for targeted intervention in this area.

[1]Since Level 0 showed complete consistency (all tested instances were appraisal elements), our data analysis begins from Level 1.

3.2.3 Annotation Pattern Analysis

A Sankey diagram in Figure 3 visualizes the annotation flow patterns and annotators' classification strategies. Annotator 1 demonstrates a conservative approach (267 annotations), preferring Propriety over Capacity (81 vs 52) within Judgement, Composition over Valuation (73 vs 40) within Appreciation. Annotator 2 exhibits a liberal approach (405 annotations), identifying more instances across categories, particularly in Reaction (51 vs 37) and Valuation (72 vs 40).

The diagram helps to identify persistent confusion areas. Within Judgement, significant cross-flows between Capacity and Propriety highlight the difficulties distinguishing the nuances of evaluative language. Appreciation subcategories reveal a significant confusion between Composition and Valuation. Furthermore, cross-category flows between Judgement and Appreciation evidence the theoretical challenges in the determination and, indeed, the ambiguation in relation to the two primary categories of Attitude. These empirical revelations and findings jointly serve to lay an informed foundation for the need of a targeted guideline refinement to address the operational disagreements in annotation scopes, category classes, and subcategory classes.

3.3 Guideline Refinement and Problem Resolution

Based on the findings arising from the pilot study, three major problematic areas were found to require targeted intervention: annotation scope clarification, Judgement versus Appreciation distinction, and subcategory classification refinement.[2] Each area received targeted treatment through specific guideline modifications and focused training exercises.

3.3.1 Annotation Scope Clarification

The pilot study revealed significant disagreements about what constitutes an appraisal expression. The differences were reviewed with a primary focus on annotations marked as 'Annotator 1 Only' and 'Annotator 2 Only', which helped to identify two major areas of disagreement:

Area 1: Factual and administrative reporting
This category includes routine procedural language that one annotator mistakenly identified as evaluative. Consider

[2] *On the one hand, Chinese state media have* ***reported*** *test kit shortages and processing bottlenecks, which could produce an undercount.* (USN0220)

In Example [2], one annotator considered "reported" as indicating a capacity, but we determined that expressions like "confirmed," "reported," and "declared" should not be annotated as they represent a procedural documentation rather than evaluation.

Area 2: Scientific terminology
Similarly, research reporting language was frequently misidentified as appraisal. Consider Example [3]:

[3] *The researchers also* ***found*** *that one deer with Omicron already had a high level of antibodies.* (USN0222)

Here, scientific reporting verbs including "reveal," "found," and "decide" in research contexts function as neutral technical descriptors without evaluative implications.

The two areas jointly constituted a disproportionately large portion of our news corpus and addressing them resulted in significant improvement.

[2]It should be noted that throughout the multiple rounds of training, the annotators encountered several other minor issues, however, this study primarily focuses on addressing the three most significant problems identified during the pilot phase.

3.3.2 Judgement versus Appreciation Distinction

The second major problematic area involved confusion between Judgement and Appreciation, as evidenced by the cross-category flows in Figure **??**. This theoretical challenge has been widely recognized with various proposed solutions. Unlike Bednarek (2009)'s non-prioritized dual-criteria approach, Taboada and Carretero (2012)'s lexical-priority ethics-aesthetics distinction, and Starfield et al. (2015)'s subcategory relocation strategy that over-broadens Valuation, our refined framework builds upon Martin and White (2005) and Thompson (2014) but goes further by establishing systematic target-based classification through entity subdivision and dual-test verification for complex cases.

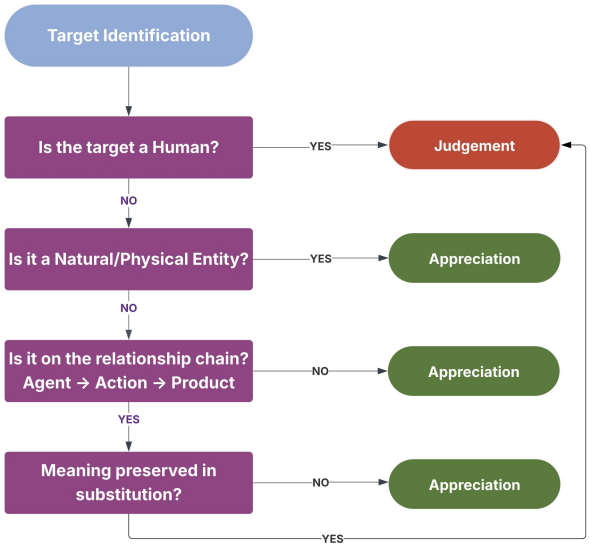

Figure 4: Decision tree for Judgement-vs-Appreciation classification

As illustrated in Figure 4, our framework begins with target identification. By establishing what entity is being evaluated as the foundational step, it ensures that classification decisions are grounded in the evaluative relationship rather than solely in lexis, thereby changing the traditionally lexical orientation of evaluative analysis. We first distinguish between human and non-human targets as the most fundamental step. Human targets receive Judgement classification, consistent with Martin and White (2005). In practice, our approach extends this to include organizations, institutions, and government entities, recognizing that these entities function as collective human agents, as illustrated in the following example:

[4] *Lipkin said he knew of one lab running 5,000*

samples a day, which might produce some false-positive results, [inflating]$_{Judgement}$ *the count.* (USN0220)

In Example [4], the appraised entity "lab" is treated as a collective human agent, and "inflating" is then classified as Judgement.

For non-human targets, rather than directly assigning all non-human targets to Appreciation, an approach that appears to simplify Judgement-vs-Appreciation distinction but creates complications in subsequent subcategorization, we subdivide non-human targets into Natural/Physical Entities and Human-Derived Entities. Natural/Physical Entities receive direct Appreciation classification, consider this example:

[5] *Omicron immediately [caused concern]*$_{Appreciation}$ *in the scientific community because it had 50 mutations compared with the original virus, many of which were known to produce.* (UKG0122)

In [5], the evaluative segment "caused concern" has "Omicron" as its appraised entity, which is a natural/physical entity and therefore classified as Appreciation.

Human-Derived Entities undergo two tests for Judgement classification:

Test 1: Agent → Action → Product relationship: Does the entity represent a human-action product traceable to human agents?

Test 2: Substitution test for meaning preservation: Does evaluative meaning remain consistent when transferred from product to agent?

We demonstrate these criteria through the following examples.

[6] *That development threatens what had been one of the most [important]*$_{Appreciation}$ *defenses against Covid: monoclonal antibodies.* (USN1122)

In [6], while "defenses" derives from human action, transferring the evaluation from "important defenses" to "the developers are important" changes the semantic meaning, failing the substitution test. Consequently, it receives Appreciation classification.

In contrast, consider an example that passes both tests:

[7] *Some information was [fabricated]*$_{Judgement}$ *to spread panic on purpose.* (CNC0620)

Example [7] passes both tests. In this example, "fabricated information" can be traced to human agents (those who fabricated it) and the evaluation transfers meaningfully to the agents: "the agents fabricated information" preserves the evaluative meaning, warranting Judgement classification.

This approach aligns with Martin and White (2005)'s and Thompson (2014)'s emphasis on target-based classification while maintaining practical clarity. By establishing systematic criteria for complex cases, our framework retains theoretical consistency while providing clear decision-making procedures for the category ambiguities that generated substantial disagreements in the pilot study.

3.3.3 Subcategory Classification Refinement

The third major problematic area emerged from fine-grained confusions within the main categories, particularly among the subcategories of Judgement, with subcategory agreement declining to F_1 = 0.460 and Cohen's Kappa values showing moderate agreement (κ = 0.360 overall, κ = 0.429 for Judgement subcategories). The Sankey diagram in Figure 3 demonstrated substantial cross-flows between related subcategories, indicating recurring confusion in fine-grained distinctions.

To address these issues, we developed an interactive clarification tool for annotators encountering classification difficulties during the annotation. This tool provides detailed distinctions for confusable pairs of subcategories, emphasizing functional rather than purely lexical distinctions, with actual annotation examples.

This reference tool allows annotators to quickly resolve subcategory uncertainties during the annotation process, directly targeting the confusion patterns revealed in our analysis of the pilot study. It offers a comprehensive facility for the fine-grained classification decisions that prove most challenging in the initial pilot phase.

4 Results and Discussion

4.1 Progressive Training and Assessment Results

Following targeted guideline refinement, we implemented iterative training to address identified problems. Training began with Round 1 (initial pilot study) that revealed three major problems. After implementing guideline refinements addressing annotation scope, Judgement-vs-Appreciation boundaries, and subcategory distinctions, we conducted successive rounds with targeted feedback and problem-specific interventions. Rounds 2-3

focused on applying refined guidelines, with Round 4 providing final assessment.

As shown in Figure 5, progressive training markedly increased annotation consistency across rounds. Subcategory-level agreement rose from moderate (κ = 0.360) to strong (κ = 0.794), while match rates improved consistently, indicating annotators achieved consistent boundary identification and scope determination. Final assessment reached 85.5% for Main Category agreement and 79.4% for subcategory classification, demonstrating that reliable inter-annotator agreement for fine-grained evaluative annotation can be achieved through progressive training and refined guidelines. These empirical results have provided a convincing validation of our approach and methodology, establishing a replicable framework for achieving the consistent annotation of evaluative language while maintaining a theoretical consistency with the established principles of AT.

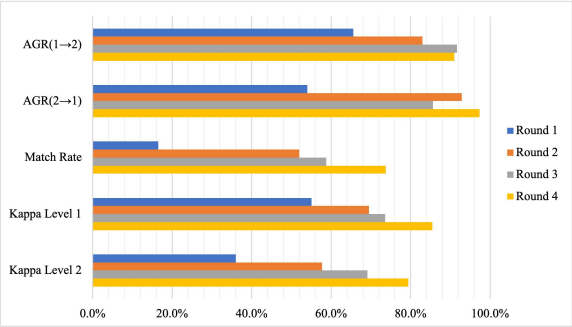

Figure 5: Consistency metrics across four progressive training rounds

4.2 Theoretical and Practical Contributions

This study has significantly contributed to the annotation of evaluative language and related methodologies in several key areas. Firstly, it has established a replicable progressive training workflow capable of transforming novice annotators through structured theoretical instruction and iterative practical feedback. Secondly, our hybrid annotation and comparison tool has proved to be effective in addressing the critical gap in existing annotation software systems, particularly for its detailed intermediate comparison and review process. Thirdly, we introduce a six-category pre-classification system that enhances accuracy and clarity in annotation comparisons, enabling precise targeting of quality control interventions. Finally, the theoretical refinements proposed in this study, particularly regarding Judgement-vs-Appreciation boundary determinations and treatment of human-derived entities, have offered practical and effective operational guidelines while maintaining an alignment with Martin and White (2005)'s foundational principles.

5 Conclusion

This article has described a precursor study in preparation for an international standard (*ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16*) for the annotation of evaluative language, aiming at providing an empirical basis for the setting of the standard. The study addressed a critical gap in AT research by developing a systematic annotation workflow that achieved reliable inter-annotator agreement for fine-grained analysis of evaluative language that is inherently subjective and ambiguous. Through progressive training protocols and targeted guideline refinements, the study yielded substantial improvements in reliability and consistency, progressing from a moderate agreement (κ = 0.360) in an initial pilot study to a strong consistency (κ = 0.794) in the final assessment. The research reported in this article has contributed to methodological issues relating to the application of AT in three key areas: a replicable workflow to ensure the reliable annotation of evaluative language, the refined theoretical guideline to address persistent classification challenges found particularly in Judgement-vs-Appreciation distinction, and a tool-facilitated approach to ensure targeted corrections and quality control.

Several limitations warrant further research. The study focused on news discourse within the COVID-19 domain, potentially limiting generalizability to other text types and evaluative contexts. The two-annotator design, while sufficient for establishing methodological principles, would benefit from an expansion to multiple annotators for broader validations. Additionally, the theoretical refinements require further testing across diverse discourse domains to confirm their general applicability. Future research should also explore human-AI collaborative annotation approaches, where our refined guidelines and training protocols could inform AI model development, while AI-assisted pre-annotation could potentially enhance human annotator efficiency and consistency in evaluative language annotation.

Acknowledgement

Research described in this article was partially supported by grants received from China National Social Science Fund (Project No 24&ZD28), City University of Hong Kong (Project Nos 9361013, 7020036 and 9360115) and Beijing Social Sciences Foundation (Project Nos 18JDYYA005 and 19YYA001).

References

Monika Bednarek. 2009. Language patterns and ATTITUDE. *Functions of Language*, 16(2):165–192.

Nancy Chinchor. 1998. MUC-7 test scores introduction. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Matteo Fuoli. 2018. A stepwise method for annotating APPRAISAL. *Functions of Language*, 25(2):229–258.

J. R. Martin and P. R. White. 2005. *The language of evaluation: Appraisal in English*. Palgrave Macmillan, Basingstoke.

Mick O'Donnell. 2018. UAM corpus tool. Computer software, version 3.3.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in English. *Language Resources and Evaluation*, 46(3):421–447.

Sue Starfield, Bridget Paltridge, Robert McMurtrie, Anthony Holbrook, Allyson Bourke, Susan Fairbairn, and Toni Lovat. 2015. Understanding the language of evaluation in examiners' reports on doctoral theses. *Linguistics and Education*, 31:130–144.

Maite Taboada and Mar´ıa Carretero. 2012. Contrastive analyses of evaluation in text: Appraisal in English, German, Spanish and French. *Linguistics and the Human Sciences*, 6(1-3):275–295.

Geoff Thompson. 2014. AFFECT and emotion, target-value mismatches, and Russian dolls: Refining the APPRAISAL model. In Geoff Thompson and Laura Alba-Juez, editors, *Evaluation in Context*, pages 47–66. John Benjamins, Amsterdam.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.