Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ontology-Guided Citation Intent Classification for Interpretable Scholarly Information Discovery

Muhammad Bilal Rehman

bilal.rehman@numl.edu.pk

National University of Modern Languages

Momina Mir

Institute of Space Technology

Shakir Ullah

Harbin Engineering University

Research Article

Keywords: Citation intent classification, Citation function, Interpretable machine learning, Ontology-driven text classification, Scientometrics, Scholarly knowledge discovery

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10848721/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ontology-Guided Citation Intent Classification for Interpretable Scholarly Information Discovery

Muhammad Bilal Rehman[1*], Momina Mir[2] and Shakir Ullah[3]

[1*]Department of Software Engineering, National University of Modern Languages, Islamabad, Pakistan.
[2]Department of Electrical Engineering, Institute of Space Technology, Islamabad, Pakistan.
[3]Department of Power Engineering and Engineering Thermophysics, Harbin Engineering University, Harbin, China.

*Corresponding author(s). E-mail(s): bilal.rehman@numl.edu.pk;

Abstract

Citation intent describes how scholarly work is used, extended, compared, or criticized, but citation counts do not capture these relationships. This study presents an ontology-guided framework that combines explicit citation-semantic features with contextual language representations to make citation-intent classification inspectable: part-of-speech-specific linguistic features, TF-IDF, SciBERT embeddings, and signals from the Citation Context and Reason Ontology (CCRO), which defines eight citation-intent categories. We evaluate it on 8,736 manually annotated ACL Anthology Network citation sentences using stratified five-fold cross-validation over five seeds, with macro-averaged F1 as the primary metric. Under matched conditions without imbalance handling it reaches a macro-F1 of 0.548 against 0.531 for fine-tuned SciBERT; the 1.7-point difference is not statistically significant, indicating comparable predictive performance rather than an accuracy advantage. Under training-fold oversampling, ablation shows the ontology-derived representation to be the largest marginal contributor among the four streams, reducing ensemble macro-F1 from 61.80% to 58.25% when removed. Because the trigger dictionary and the annotation scheme share a conceptual taxonomy, this measures an explicit, inspectable prior rather than independent ontological evidence. An instance-level audit trail links each prediction to its activated ontology category, matched lexical evidence, and the class predicted when the ontology stream is withheld. Structured citation knowledge can therefore be combined with contextual representations to support transparent scholarly information analysis without sacrificing predictive parity.

Keywords: Citation intent classification, Citation function, Interpretable machine learning, Ontology-driven text classification, Scientometrics, Scholarly knowledge discovery

1 Introduction

Understanding why a scholarly work cites another is important for interpreting how scientific knowledge is developed, extended, compared, and criticized [1]. Conventional citation counts measure the volume of scholarly attention but do not distinguish the rhetorical or functional role of individual citations [1]. Two papers receiving the same number of citations may therefore occupy different positions in a knowledge network: one may be used as a methodological foundation, another may be criticized, and another may be mentioned as background. Advances in natural language processing now make it possible to classify the function of a citation directly from its surrounding text, rather than only counting [2]. Computational identification of citation intent can complement quantitative citation indicators by providing a semantic layer for scholarly information analysis.

Computational approaches have moved from manually engineered linguistic features to contextual neural representations. Early methods combined hand-crafted linguistic features with conventional classifiers [3], but these features do not separate closely related intents well. Traditional classifiers using cue phrases, lexical features, reporting verbs, and statistical representations provide relatively transparent decision signals but have difficulty distinguishing semantically related citation functions. Pretrained scientific language models such as SciBERT [4] provide dense contextual representations and perform well on benchmark datasets [5], but they still do not distinguish neighboring cognitive intents, for example whether a method is incorporated or only compared. More recently, large language models have been applied to the task, but their behavior is inconsistent and hard to reproduce across systems, and the highest scores tend to rely on opaque, proprietary models [6]. Ensemble and prompt-based pipelines raise accuracy further [7, 8], usually at the cost of transparency. These developments have largely emphasized predictive performance, while the semantic evidence underlying individual classification decisions remains difficult to inspect systematically.

A practical gap therefore remains between predictive citation-intent classification and the transparent representation of the evidence supporting individual decisions. Existing neural representations provide contextual information but do not expose domain-level citation semantics. Ontology-based approaches provide explicit semantic categories but lack the contextual sensitivity of neural representations [9]. This raises the question of whether structured citation knowledge can complement contextual representations while retaining an inspectable decision pathway.

We investigate an ontology-guided representation that combines explicit citation semantics with contextual and lexical representations. The objective is to determine whether ontology-derived features provide complementary information and enable more transparent analysis of classification behaviour, rather than to assume that they

improve predictive accuracy. We add two lexical resources, WordNet and VerbNet, and the domain knowledge in the Citation Context and Reason Ontology (CCRO) [10] to the feature space. The resulting ontological signal is an auditable component of the classifier, so each prediction traces back to identifiable linguistic and ontological evidence. The framework combines part-of-speech (POS) specific lexical features, term frequency-inverse document frequency (TF-IDF) statistics, SciBERT embeddings, and CCRO-derived features for eight-class classification. We evaluate it against a statistical baseline, a standalone ontology matcher, a fine-tuned SciBERT, and a few-shot large language model under stratified five-fold cross-validation, and report macro-averaged F1 because of the severe class imbalance in the corpus. Imbalance handling is matched across systems so that the contribution of the representation is separated from the contribution of oversampling. We characterise the trigger lexicon by category coverage, and evaluate the per-instance audit trail that the architecture produces. We treat the ontology-derived representation as an explicit, human-auditable prior rather than as a source of independent predictive power, and we measure both its marginal contribution and the faithfulness of the resulting decisions.

This study addresses the following research questions:

1. To what extent can ontology-derived features complement lexical, statistical, and contextual representations for fine-grained citation intent classification?
2. What is the marginal contribution of ontology-derived features relative to other representation streams?
3. How do ontology-guided representations affect the interpretability and diagnosis of citation-intent classification errors?

The contributions of this study are threefold. First, we develop an ontology-guided citation-intent classification framework that combines contextual language representations with explicit citation-semantic features. Second, we show through matched comparison and stream ablation that the ontology-derived stream contributes complementary signal while predictive performance relative to a fine-tuned neural baseline is preserved. Third, we instantiate a per-instance audit trail that links each prediction to its activated ontology category and matched lexical evidence, so the role of the ontological prior is inspected rather than inferred from a post-hoc explanation.

The rest of the paper reviews related work, presents the CCRO-based framework and its feature extraction, reports the experimental setup and results, and discusses predictive performance, interpretability, scholarly-information implications, and limitations.

2 Related work

2.1 Citation intent and scholarly communication

Research on citation intent predates computational text analysis. Moravcsik and Murugesan [11] gave an early systematic account, classifying references along functional dimensions: conceptual versus operational, organic versus perfunctory, and confirmative versus negational. They found that many citations serve rhetorical or procedural purposes rather than marking direct intellectual influence. Chubin and

Table 1 Position of the proposed approach relative to existing lines of work.

Approach	Contextual representation	Explicit citation semantics	Fine-grained	Interpretability
Traditional linguistic models	Limited	Limited	Yes	High
SciBERT-based models	High	No	Yes	Limited
LLM/prompt approaches	High	No	Yes	Limited
Ontology-based approaches	Limited	Yes	Yes	High
Proposed approach	High	Yes	Yes	Explicitly investigated

Moitra [12] proposed a hierarchical framework around the same time, distinguishing supportive and critical citation. These studies questioned whether citation counts alone measure scholarly impact and framed later work on the semantic and rhetorical functions of citations. The limits of purely quantitative bibliometric indicators are still recognised in judgements of research influence and knowledge transfer [1].

2.2 Computational citation-intent classification

Automation began at the turn of the century. Nanba and Okumura [13] used citation-relationship types for multi-paper summarisation. Garzone and Mercer [14] built an early rule-based classifier over a multi-category scheme. Nakov, Schwartz, and Hearst [15] introduced *citances*, citation sentences treated as the unit of semantic analysis, which fixed the citation sentence as the object of study for later function classification, including ours. Teufel, Siddharthan, and Tidhar [16] scaled classification with a twelve-category citation function taxonomy grounded in argumentative zoning, treating citation intent as a rhetorical relationship within a discourse context rather than a fixed property of the cited work. Abu-Jbara, Ezra, and Radev [17] extended this with the ACL-ARC framework, categorising citations by purpose and polarity through functions such as critique, comparison, use, substantiation, and basis. ACL-ARC is now a common benchmark. Jurgens et al. [18] built a large annotated corpus from the ACL Anthology with a citation frame taxonomy covering background, motivation, use, extension, comparison, and future work, and showed how citation behaviour shifts over time within a discipline.

Before deep learning, classification relied mainly on supervised methods with hand-engineered linguistic features. Roman et al. [2] showed that lexical representations from citation contexts could separate citation functions with traditional classifiers. Support vector machines, random forests, cue phrases, reporting verbs, and n-gram features were competitive on benchmark datasets, which suggests that citation intent often surfaces in local linguistic patterns [3]. WordNet-based synonym and hypernym relations moved these representations past surface form [19], and VerbNet mappings added verb-class information for detecting methodological adoption and extension [20]. TF-IDF weighting sharpened feature discrimination by favouring informative terms [21]. These methods were interpretable and cheap to run, but their reliance on surface signals made closely related citation functions hard to separate.

2.3 Ontologies and semantic representations

Ontology-driven approaches formalise citation semantics through structured knowledge. Peroni and Shotton [22] introduced the Citation Typing Ontology (CiTO) within the SPAR suite, an OWL-encoded framework for describing citation relationships and their rhetorical characteristics in semantic publishing. Ihsan and Qadir [10] built a more specialised ontology, the Citation's Context and Reasons Ontology (CCRO), which consolidates a large set of citation reasons into eight disjoint categories supporting both sentiment-oriented and rhetorical readings. Domain-specific ontologies can impose explicit semantic constraints on otherwise unconstrained pipelines, confirmed in text classification tasks that require interpretability [9]. This matters in scientometric work, where the transparency of a classification decision affects the credibility of the research evaluation built on it.

2.4 Neural and large-language-model approaches

Deep neural architectures improved classification by learning contextual representations directly. Cohan et al. [5] proposed a multitask framework with auxiliary objectives for document structure and citation worthiness, which outperformed feature-engineered approaches and produced the SciCite dataset, a standard cross-domain evaluation resource. SciBERT [4], a transformer pre-trained on scientific publications, outperformed general-domain models on scientific NLP tasks through domain-specific vocabulary and context. Lauscher et al. [23] questioned the assumption that a citation can be captured by a single sentence and a single label. Their Multi-Cite dataset showed that citation intent often spans several sentences and expresses more than one function at once, which single-label schemes cannot represent.

Multi-task learning has continued to improve results. Zhang et al. [24] showed that optimising citation intent classification with related objectives gives measurable gains on benchmark datasets. Large language models raise the option of citation intent classification through prompting and few-shot learning with little task-specific annotation [6, 8], but these methods often drop the explicit reasoning that ontology-based frameworks keep, and their results have been hard to reproduce across systems [6].

2.5 Research gap

For scholarly knowledge graph construction, research evaluation, and scientometric discovery, transparent classification decisions remain essential [25]. A meta-analysis of semantic citation classification [26] sets the context against which single-system results, including ours, should be read. This points to the hybridisation hypothesis examined here: combining contextual language representations with structured semantic knowledge to balance predictive performance against interpretability, and to ground high-dimensional neural embeddings in human-interpretable semantic structures.

Table 1 shows the position of the present study: rather than proposing another contextual classifier, we investigate whether explicit citation semantics can be incorporated into a contextual classification pipeline in a way that supports both prediction and diagnosis.

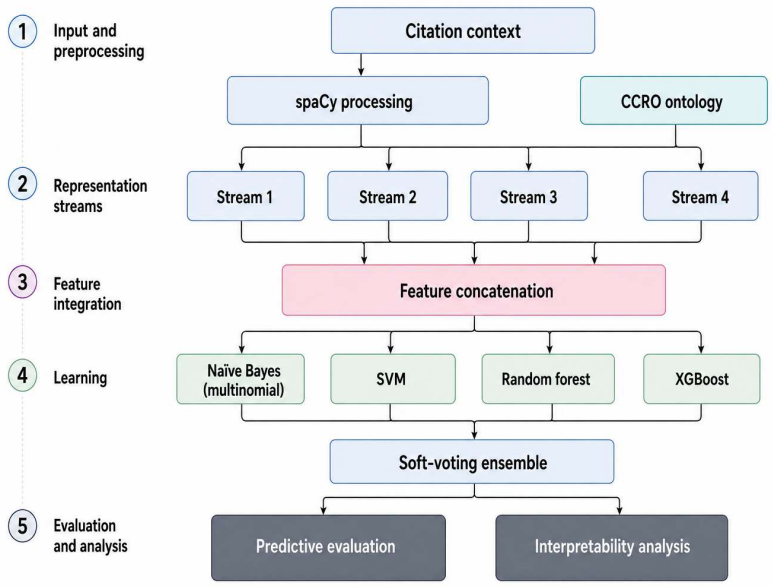

Fig. 1 Proposed methodology pipeline for citation intent classification.

3 Methodology

3.1 Research framework

The proposed framework classifies citation intent by combining structured ontological knowledge with statistical and neural feature representations. Figure 1 shows the full pipeline, from raw citation text in the ACL Anthology Network (AAN) corpus to the final eight-class output.

Every citation sentence passes through the same preprocessing pipeline before feature extraction. The spaCy English model[27] tokenises, lemmatises, and POS-tags each sentence. On a stratified random sample of 200 AAN citation sentences, spaCy reaches a validated POS-tagging accuracy of about 89.7%. This is below its accuracy on general English, which we attribute to the density of technical terminology, inline mathematical notation, and non-standard orthography in scientific citation contexts. Tokens are then split into grammatical subsets (nouns, verbs, adjectives, adverbs) for POS-specific extraction. The pipeline runs identically across all streams, so the four representations are computed from the same normalised input.

3.1.1 Methodological rationale.

The four streams were selected because they encode different kinds of evidence, not to maximise the number of input features. Linguistic-semantic features represent the lexical and grammatical patterns through which citation intent is expressed [3]; TF-IDF captures corpus-level lexical discriminativeness; SciBERT supplies a contextual

Table 2 CCRO sentiment classes with descriptions.

Polarity	Class	Description
Positive	Incorporate	Cites the prior work as a component integrated into the current study.
Positive	Extend	Builds upon prior work to reach a broader solution.
Positive	Based on	Uses prior work as the foundational starting point.
Negative	Negate	Argues that prior work is invalid or ineffective.
Negative	Criticize	Identifies specific weaknesses or faults in prior work.
Negative	Contrast	Highlights differences between prior work and the current approach.
Neutral	Compare	Draws similarities between prior work and the current study.
Neutral	Discuss	References prior work in a general descriptive manner.

representation of the citation sentence [4, 5]; and CCRO supplies an explicit ontology-derived representation of citation reasons [9, 10]. The streams therefore operate at different representational levels.

Concatenation was adopted as the integration mechanism because it keeps each stream identifiable and so allows its marginal contribution to be measured by removal (Section 4.5). The combined representation is not assumed to be optimal. Whether each stream carries information the others do not is treated as an empirical question, tested through matched baselines (Section 4.4), the standalone and pairwise ontology experiments (Section 4.3.1), stream ablation (Section 4.5), and instance-level analysis (Section 4.6).

3.2 CCRO representation of citation intent

The framework uses the Citation Context and Reason Ontology (CCRO),[10] a formal semantic structure for the cognitive relationships expressed in citation contexts. Each citation sentence c_i is assigned to one of eight ontological classes in $C = \{c_1, c_2, \ldots, c_8\}$ through the mapping $f_{\text{CCRO}} : T \to C$, where T is the preprocessed citation text. The taxonomy is

$$\begin{aligned} \text{Class}(c_i) \in \{&\text{Incorporate},\ \text{Extend},\ \text{Based on},\\ &\text{Negate},\ \text{Criticize},\ \text{Contrast},\\ &\text{Compare},\ \text{Discuss}\}. \end{aligned} \tag{1}$$

The eight classes are grouped into three sentiment polarities (positive, negative, and neutral) by the rhetorical role of the citation with respect to the cited work. Table 2 gives the full definitions.

The ontology is used as a semantic representation rather than as a replacement for contextual language modelling. Each citation context may activate an explicit category-level prior while the neural representation retains information from the broader linguistic context. This distinction matters because lexical triggers do not

uniquely determine citation intent. The ontology-derived representation is therefore treated as evidence available to the classifier rather than as a deterministic label assignment.

Each citation context is characterised along three dimensions: sentiment polarity $S \in \{-1, 0, 1\}$, rhetorical background R, and methodological relationship M between the citing and cited papers. The composite representation is

$$\psi(c_i) = \langle S(c_i),\ R(c_i),\ M(c_i) \rangle. \tag{2}$$

Here ψ is the three-dimensional characterisation of the ontology. It is distinct from the eight-dimensional feature vector $\phi(T)$ of Section 3.6, which is the quantity supplied to the classifier.

3.3 Linguistic representation (Stream 1)

For each token t_j in T, we extract verb classes $V(t_j)$ from VerbNet,[28] noun, adjective, and adverb synsets $W(t_j)$ from WordNet,[29] and the POS tag sequence $P(T)$. Token-level resources are pooled to sentence level by bag-of-identifiers aggregation: the VerbNet class and WordNet synset identifiers activated anywhere in the sentence are collected into binary indicator vectors, and the POS tag sequence is encoded as tag n-gram counts. The three components are concatenated into a sentence-level block

$$L(T) = \Big[\bigvee_j V(t_j) \parallel \bigvee_j W(t_j) \parallel P(T) \Big], \tag{3}$$

where $\bigvee$ is element-wise disjunction over the tokens of T. These resources let the model separate, for example, verbs of methodological incorporation (*builds upon*, *adopts*) from verbs of critical contrast (*contradicts*, *differs from*).

3.4 Statistical representation (Stream 2)

The TF-IDF weight for each term t in citation document d is

$$w_{(t,d)} = \mathrm{tf}_{(t,d)} \times \log\left(\frac{D}{\mathrm{df}_t}\right), \tag{4}$$

where D is the number of documents and df_t the number containing term t. This emphasises terms informative to a specific citation context and down-weights terms that appear uniformly across the corpus. TF-IDF is retained alongside Stream 3 because the two encode different information: TF-IDF is a sparse representation of corpus-level lexical discriminativeness, whereas SciBERT is a dense representation of sentence context. Keeping both allows the marginal contribution of sparse lexical statistics to be measured separately (Section 4.5).

3.5 Contextual representation (Stream 3)

Sentence representations come from SciBERT,[4] a transformer pre-trained on scientific text. The embedding $E(T)$ is taken from the `[CLS]` token:

$$E(T) = \text{SciBERT}(T)_{[\text{CLS}]}. \tag{5}$$

This 768-dimensional vector captures long-range semantic dependencies and domain-specific patterns that surface lexical features cannot reliably detect.

3.6 Ontology-derived representation (Stream 4)

To encode domain knowledge, CCRO membership is determined using a rule-based matcher that scans each preprocessed sentence for lexical triggers from a dictionary of 194 entries spanning all eight categories, for example *builds upon* (Extend) and *differs from* (Contrast). Each sentence is represented as an eight-dimensional indicator vector $\phi(T)$ describing which categories, if any, the trigger matching activated.

The dictionary was constructed from the semantic definitions and linguistic analysis underlying CCRO rather than by fitting the dictionary to the evaluation labels. The matcher runs on raw text without access to ground-truth annotations, so its construction is independent of the labelled data used for evaluation.

Because the annotation scheme and CCRO originate from the same conceptual taxonomy, the triggers and gold labels share a conceptual source. Stream 4 is therefore an *ontology-informed prior* of the weak-supervision kind: an explicit, human-auditable hypothesis about each sentence's category, derived from the taxonomy that guided annotation rather than from evidence outside it. The classifier is free to override the prior, and the experiments measure how often it does so.

Three operational consequences follow. The matcher sees only raw text. Its output is an eight-dimensional indicator vector that the downstream ensemble may use or ignore. Its contribution is assessed only through the standalone and pairwise baselines (Section 4.3.1), stream ablation (Section 4.5), and the per-instance audit trail (Section 4.6), not through any claim that the ontology supplies knowledge independent of the annotation scheme.

3.7 Feature fusion

The four representation streams are concatenated to form a unified feature vector v:

$$v = \big[L(T) \parallel w_{(\cdot,d)} \parallel E(T) \parallel \phi(T)\big], \tag{6}$$

where $\parallel$ denotes vector concatenation, $L(T)$ is the pooled Stream 1 block from Equation (3), and $w_{(\cdot,d)}$ is the TF-IDF vector for document d. The representation combines structured linguistic knowledge (Streams 1 and 4), the distributional statistical signal (Stream 2), and the contextual neural representation (Stream 3).

Concatenation preserves the identity of each information source and enables controlled stream-level ablation. The unified vector v maintains the separability of the four streams: the pooled syntactic structure $L(T)$, the document-level term weights

$w_{(\cdot,d)}$, the contextualised embeddings $E(T)$, and the ontology-derived indicators $\phi(T)$. Because the objective is to examine complementarity among these information sources, separable streams support the analysis of marginal contributions (Section 4.5) and the stream-removal analysis of Section 4.6 without imposing learned interactions that would obscure source-specific effects.

3.8 Classification and ensemble strategy

The vector v is passed to four supervised classifiers trained in parallel: multinomial Naïve Bayes with Laplace smoothing, an SVM with an RBF kernel and class-weighted loss, a random forest [30] of 100 decision trees with entropy criterion, and XGBoost [31] with early stopping, all in scikit-learn [32]. The four were chosen to span heterogeneous inductive biases rather than to identify a single best classifier: a generative probabilistic model, a kernel method, a bagged tree ensemble, and a boosted tree ensemble. Evaluating the representation across these families reduces the chance that the observed behaviour is specific to one learning algorithm. Each component classifier produces a per-class probability distribution over the eight categories. The final prediction is obtained via soft voting, in which the four distributions are averaged element-wise before selecting the class with the highest aggregate probability. Soft voting retains the confidence information produced by each classifier, avoids deterministic tie resolution based on class-index ordering, and supplies the per-instance confidence values used in the interpretability analysis (Section 4.6). The ensemble is a secondary aggregation step rather than the methodological contribution of the study; its purpose is to test whether combining heterogeneous classifiers gives more stable performance than any single one.

Multinomial Naïve Bayes requires non-negative inputs, but the SciBERT `[CLS]` embeddings contain negative components. The fused vector is min-max scaled to the unit interval before this member is fitted. This makes the multinomial likelihood well defined and is why the smoothing parameter α is the quantity tuned for it in Table 5. The scaled features are continuous values rather than the term counts the model assumes, so its generative assumptions hold only approximately. The large smoothing value chosen for the All-POS subset ($\alpha = 6$) is consistent with this member contributing a flat, prior-dominated distribution.

3.8.1 Imbalance handling.

The AAN corpus is strongly skewed toward the Discuss category (Section 4.1), so any reported difference between two systems may reflect imbalance treatment as much as representation. We therefore apply two balancing strategies and hold them constant across the compared systems (Section 4.4).

The first is class-weighted cross-entropy, with weights set inversely proportional to class frequencies in the training folds. This changes the loss but leaves the feature distribution unaltered.

The second is SMOTE [33], applied only within the training folds so that each test fold retains the natural class distribution. Two properties of SMOTE on this feature space should be recorded. It interpolates linearly between neighbouring instances, so

on the concatenated vector v it produces fractional values in the ontological block $\phi(T)$: synthetic instances whose CCRO activation state the trigger matcher could not emit. It also loses effectiveness in the high-dimensional regime the TF-IDF block occupies, where near-neighbour interpolation becomes unreliable [34]. Class weighting is therefore the primary balancing condition for the proposed framework, and SMOTE results are reported alongside it. A categorical-aware variant such as SMOTE-NC [33] or a focal-loss alternative would be a more direct treatment of the mixed feature space.

3.9 Interpretability mechanism

Because the streams remain separable in v, the framework produces a per-instance record of the evidence available to the classifier. For each citation context the system stores: the activated CCRO category or categories from the trigger matcher together with the specific triggers that fired; the VerbNet classes and WordNet synsets activated in Stream 1; the highest-weighted TF-IDF terms in Stream 2; the predicted class and soft-voting confidence from the full model; and the predicted class and confidence obtained when an individual stream is removed and the model retrained on the same training folds. This record is the audit trail evaluated in Section 4.6.

Two quantities are defined over this record. The first, *interpretability consistency*, is the proportion of instances for which the final prediction agrees with the ontology-derived category carrying the strongest trigger activation. Let $a(c_i)$ be the class with the highest activation in $\phi(T)$ for instance i and $p(c_i)$ the final prediction. Instance i is consistent if $p(c_i) = a(c_i)$. Instances for which no trigger fires, that is all entries of $\phi(T)$ are zero, carry no ontological signal and are excluded from the calculation. The measure addresses faithfulness: it records how often the decision follows the explicit, human-auditable evidence rather than overriding it.

The second, *stream sensitivity*, is the change in prediction confidence when a stream is removed. Let $q(c_i)$ be the maximum soft-voting probability from the full model and $q^{-k}(c_i)$ the corresponding value from the model retrained without stream k. The sensitivity of stream k on instance i is

$$\Delta_k(c_i) = q(c_i) - q^{-k}(c_i). \tag{7}$$

A positive Δ_k indicates that stream k raises the confidence assigned to the predicted class. We report the mean Δ_k over the sampled instances together with the proportion of instances whose predicted class changes when stream k is removed. Stream 3 serves as the comparison, being the largest and most information-dense of the four.

3.10 Evaluation protocol

3.10.1 Metrics.

The primary metric is macro-averaged F1, which weights each class equally and suits the imbalanced distribution:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k, \tag{8}$$

with $K = 8$ on the AAN corpus. Precision, recall, and F1 per class are

$$P = \frac{TP}{TP + FP}, \tag{9}$$

$$R = \frac{TP}{TP + FN}, \tag{10}$$

$$F_1 = \frac{2 \times P \times R}{P + R}, \tag{11}$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative predictions. Weighted F1 is reported alongside macro-F1 to allow comparison with prior work that uses a different primary metric. Macro-F1 is primary throughout, because it prevents strong majority-class performance from masking weak minority-class performance on classes such as Negate and Incorporate.

3.10.2 Cross-validation and seeds.

Model selection and evaluation use stratified five-fold cross-validation over all 8,736 sentences, preserving the class distribution within every fold. To address the seed sensitivity of SMOTE and of the stochastic components of the ensemble, the full pipeline (preprocessing, feature extraction, training-fold balancing, classifier training, and evaluation) is run with five independent random seeds, $\{42, 123, 456, 789, 101112\}$.

This yields two levels of aggregation. Each of the five folds for a given seed produces one macro-F1 estimate, giving 25 fold-level observations per condition. The mean over the five folds of a seed gives a seed-level mean, and we report the mean and standard deviation over the five seed-level means:

$$\begin{aligned} \bar{F}_s &= \frac{1}{5}\sum_{f=1}^{5} F_{s,f}, \qquad \bar{F} = \frac{1}{5}\sum_{s=1}^{5} \bar{F}_s, \\ \sigma &= \sqrt{\frac{1}{4}\sum_{s=1}^{5}(\bar{F}_s - \bar{F})^2}, \end{aligned} \tag{12}$$

where $F_{s,f}$ is the macro-F1 on fold f for seed s. All configurations are reported as $\bar{F} \pm \sigma$. The standard deviation captures variation from SMOTE synthesis, SVM optimisation initialisation, random forest bootstrap sampling, and XGBoost subsampling, and so measures the stability of the reported differences.

3.10.3 Significance testing.

The 25 fold-seed observations are not independent. The five folds within a seed partition the same 8,736 sentences, and every seed reuses that corpus, so training sets overlap heavily both within and across seeds. An ordinary paired t-test over these observations underestimates the variance of the mean difference and is therefore anti-conservative [35]. We use instead the corrected resampled t-test of Nadeau and Bengio [36], which inflates the variance term to account for the overlap.

Table 3 Research questions and the experiments that address them.

RQ	Experiment	Reported in
1	Matched comparison against fine-tuned SciBERT under identical imbalance handling	Sections 4.3, 4.4
2	Standalone ontology matcher, pairwise CCRO combinations, and leave-one-stream-out ablation	Sections 4.3.1, 4.5
3	Interpretability consistency, stream sensitivity, per-instance audit trail, and confusion analysis	Sections 4.6, 4.7

For a comparison between configurations A and B we compute the per-fold difference $d_{s,f} = F^{(A)}_{s,f} - F^{(B)}_{s,f}$ and the statistic

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{n} + \frac{n_2}{n_1}\right) s_d^2}}, \qquad \bar{d} = \frac{1}{n}\sum_{s=1}^{5}\sum_{f=1}^{5} d_{s,f}, \tag{13}$$

where $n = 25$ is the number of fold-seed observations, s_d^2 their sample variance, and n_2/n_1 the ratio of test-fold to training-fold size, equal to $1/4$ under five-fold cross-validation. The correction term replaces the $1/n$ of the ordinary paired test. At this fold ratio it raises the estimated standard error by a factor of approximately 2.7, so a corrected p-value is always larger than its uncorrected counterpart. The statistic is referred to $n - 1 = 24$ degrees of freedom.

We apply the Holm-Bonferroni correction [37] to control the family-wise error rate. The family comprises the three matched-condition comparisons of Section 4.4 (A vs D, B vs E1, C vs E2) and the four stream-removal comparisons of Section 4.5, seven tests in total. Where an analysis is exploratory and no family-wise control is applied, we state this and read the p-value as suggestive [38].

3.10.4 Mapping of research questions to experiments.

Table 3 states which experiment answers each research question.

4 Experimental setup and results

4.1 Dataset and annotation

Experiments use the AAN citation corpus,[39] which has 8,736 citation sentences from the computational linguistics literature, each with citing and cited paper identifiers and the context sentence. An added annotation layer assigned every sentence to one of the eight CCRO classes of Table 2. Three graduate researchers annotated after a two-day training session designed to align interpretation across the eight categories. All 8,736 sentences were labelled independently by all three annotators rather than divided among them, so agreement is computed over the full corpus rather than an overlap sample. The gold label was taken by majority: where two annotators agreed, that category was adopted. The gold standard is therefore a majority-of-three construct

Table 4 Eight-class distribution of the annotated AAN corpus ($n = 8{,}736$). Percentages are of the full corpus, rounded to one decimal place.

Polarity	Class	n	%
Neutral	Discuss	6,100	69.8
Neutral	Compare	627	7.2
Positive	Extend	420	4.8
Positive	Based on	250	2.9
Positive	Incorporate	159	1.8
Negative	Contrast	900	10.3
Negative	Criticize	190	2.2
Negative	Negate	90	1.0
	Total	8,736	100.0

rather than an expert adjudication, and because the gold labels come from the same three annotators whose mutual agreement κ measures, the reported κ characterises the internal consistency of the annotation process rather than accuracy against an external reference.[40]

With three annotators, inter-annotator agreement uses Fleiss' κ,[41] which generalises Cohen's two-rater statistic[42] to any fixed number of raters. For n items, m raters, and K categories, with n_{ik} the number of raters assigning item i to category k, the mean observed agreement $\bar{P}$ and chance agreement $\bar{P}_e$ are

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m(m-1)}\left(\sum_{k=1}^{K} n_{ik}^2 - m\right), \quad \bar{P}_e = \sum_{k=1}^{K} p_k^2, \tag{14}$$

where $p_k = \frac{1}{nm}\sum_{i=1}^{n} n_{ik}$. Fleiss' κ is then

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \tag{15}$$

Over the eight categories this gives $\kappa = 0.83$, in the "almost perfect" band of the Landis and Koch scale[43] and a demanding result given that chance agreement falls as the number of categories rises.

The dataset is imbalanced. Aggregated to polarity, neutral citations are 6,727 instances (77.0%), negative 1,180 (13.5%), and positive 829 (9.5%). The neutral polarity aggregates Discuss and Compare, so no single class reaches the 77.0% share; the imbalance is driven by one class, Discuss, at 69.8% (Table 4).

Discuss alone is 69.8% of the corpus, over five times the next largest class, and the three smallest classes (Incorporate, Criticize, Negate) together make up 5.0%. Under five-fold cross-validation each held-out fold holds roughly 18 Negate, 32 Incorporate, and 38 Criticize instances, the sample from which each per-class F1 estimate is formed.

Balancing is applied within training folds only (Section 3.8.1), leaving each test fold unmodified, so all reported metrics reflect naturally distributed data.

4.2 Baselines and comparison systems

The framework is evaluated against four comparison systems, all trained and evaluated over the same folds under identical preprocessing:

(i) a *TF-IDF + SVM* statistical baseline (TF-IDF-weighted term vectors, RBF-kernel SVM);
(ii) the *CCRO trigger matcher operating alone* as a classifier, assigning each sentence the category of its strongest lexical activation, which isolates the discriminative power of Stream 4 without any learned component;
(iii) a *fine-tuned SciBERT*[4] with a classification head (a frozen SciBERT is excluded, since without fine-tuning it is not a meaningful neural comparison);
(iv) *Qwen2.5-1.5B (few-shot)*,[44] a small open-weight model under few-shot prompting without fine-tuning.

Imbalance handling is matched between the proposed framework and the SciBERT baseline (Section 4.4), so that the contribution of the representation is separated from the contribution of the balancing strategy. Qwen2.5-1.5B is small by current standards, so we draw no general conclusion about prompt-based classification from it; larger and proprietary models present documented reproducibility difficulties on this task,[6] and our claim is restricted to the configuration tested.

CiteFusion[7] is excluded from the primary results because it uses a six-class ACL-ARC taxonomy whereas the framework uses the eight-class CCRO scheme. Since macro-F1 is the unweighted mean of per-class F1 scores, it is sensitive to the number of classes, so cross-taxonomy comparison is unsound.

4.2.1 Hyperparameter optimisation.

Hyperparameters are tuned per classifier in two phases: a random search over 50 iterations identifies promising regions, then a focused grid search refines them. The procedure is applied independently to each POS-specific subset, with the resulting configurations in Table 5. An earlier stratified 80/10/10 holdout (seed 52) served only for development hyperparameter search and underlies none of the numbers reported here.

4.3 Main results

Table 6 reports macro-F1 and weighted F1 for every comparison system alongside the proposed framework, with the difference from the fine-tuned SciBERT baseline in the final column.

Under the matched evaluation condition, the proposed representation achieves a macro-F1 of 0.548, compared with 0.531 for fine-tuned SciBERT. The difference is 1.7 percentage points and is not statistically significant. The uncorrected paired statistic gives $p = 0.053$; because the variance correction of Section 3.10.3 can only raise a p-value, the comparison does not approach significance under the corrected test.

Table 5 Optimised hyperparameters across POS-specific feature subsets.

Subset	Model	Hyperparameters
Nouns	Naïve Bayes	$\alpha = 0.5$
	SVM	kernel = rbf, $C = 18$, γ = auto
	Random forest	n_estimators = 100, criterion = entropy, max_depth = 5
	XGBoost	n_estimators = 1000, lr = 0.001, max_depth = 4, subsample = 0.8, colsample_bytree = 0.8, $\gamma = 0.1$
Adjectives	Naïve Bayes	$\alpha = 1.5$
	SVM	kernel = rbf, $C = 10$, γ = auto
	Random forest	n_estimators = 100, criterion = entropy, max_depth = 10
	XGBoost	n_estimators = 200, lr = 0.1, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8, $\gamma = 0.1$
Verbs	Naïve Bayes	$\alpha = 1.5$
	SVM	kernel = rbf, $C = 10$, γ = scale, decision_function_shape = ovo
	Random forest	n_estimators = 100, bootstrap = true, max_depth = 20, min_samples_leaf = 1
	XGBoost	n_estimators = 200, lr = 0.1, max_depth = 5, subsample = 0.8, colsample_bytree = 0.8, $\gamma = 0.1$
All POS	Naïve Bayes	$\alpha = 6$
	SVM	kernel = rbf, $C = 0.1$, $\gamma = 1$
	Random forest	n_estimators = 100, criterion = entropy, max_depth = 10
	XGBoost	n_estimators = 1000, lr = 0.01, max_depth = 8, subsample = 0.8, colsample_bytree = 0.8, $\gamma = 0.1$

Table 6 Main results on the eight-class CCRO taxonomy, averaged over five stratified cross-validation folds and five random seeds. Macro-F1 is the primary metric. Δ is the macro-F1 difference from the fine-tuned SciBERT baseline under the same balancing condition. Weighted F1 was not recorded for the few-shot baseline.

Model	Balancing	Macro-F1	Weighted F1	Δ vs SciBERT
CCRO matcher (standalone)	None	0.310	0.481	-0.221
TF-IDF + SVM	None	0.485	0.662	-0.046
Qwen2.5-1.5B	Few-shot	0.393	—	-0.138
SciBERT (fine-tuned)	None	0.531	0.724	—
Proposed	None	0.548	0.733	$+0.017$
SciBERT (fine-tuned)	Class-weighted	0.562	0.731	—
Proposed	Class-weighted	0.571	0.738	$+0.009$

Table 7 Standalone and pairwise performance of the ontological stream, without balancing. Macro-F1 is the primary metric.

Configuration	Macro-F1	Weighted F1
CCRO matcher (standalone)	0.310	0.481
TF-IDF + SVM	0.485	0.662
SciBERT (fine-tuned)	0.531	0.724
CCRO + TF-IDF	0.520	0.689
CCRO + SciBERT	0.545	0.731
Full model (four streams)	0.548	0.733

The result therefore indicates comparable predictive performance rather than a statistically demonstrated accuracy advantage. This finding matters for interpreting the contribution of the ontology-derived representation: its value should not be attributed solely to improved classification accuracy, but should instead be examined through the complementary evidence and interpretability analyses presented below.

The few-shot Qwen2.5-1.5B baseline reaches 0.393, which is not competitive on a fine-grained eight-class task.

4.3.1 Standalone and pairwise contribution of the ontological stream.

To separate the standalone discriminative power of the ontological stream from its complementarity with the statistical and neural representations, we evaluate six configurations under the same cross-validation protocol. All configurations in this subsection use no balancing, matching the condition of the SciBERT comparison above, so that the effect of the representation is observed without the effect of oversampling.

The configurations are: (1) the CCRO trigger matcher as a standalone classifier, assigning each sentence the category of its strongest lexical activation; (2) a TF-IDF + SVM baseline; (3) the fine-tuned SciBERT baseline; (4) CCRO indicators concatenated with TF-IDF features, classified by the same SVM; (5) CCRO indicators concatenated with SciBERT `[CLS]` embeddings, classified by the same neural head; and (6) the full four-stream model. Table 7 reports macro-F1 and weighted F1 for the six configurations.

The standalone matcher reaches a macro-F1 of 0.310, the lowest of the six configurations, which establishes that the lexical trigger rules are not discriminative enough to classify citation intent on their own. Their value in the framework is as a prior that the learned components refine or override, not as a classifier.

The CCRO + SciBERT configuration reaches 0.545 against 0.531 for SciBERT alone. The ontology-derived prior therefore supplies signal that the contextual representation does not already carry, which is the complementarity that the framework is built to exploit. The full model reaches 0.548, so the POS-specific and statistical streams continue to add a positive margin once the neural and ontological signals are combined.

The progression from CCRO + TF-IDF (0.520) to CCRO + SciBERT (0.545) shows that the ontological prior contributes more when paired with contextual semantics than with surface term statistics. The pattern is consistent with the CCRO features helping to resolve local lexical ambiguities, for example the separation of *Compare* from *Discuss*, that the dense embedding does not resolve alone; Section 4.6 examines this at the instance level.

4.4 Effect of imbalance handling

Imbalance handling has a substantial effect on minority-class performance on this corpus. When the same imbalance-treatment strategy is applied to both systems, the comparison provides a more controlled estimate of the contribution of the representation architecture. The results show that improvements associated with oversampling should be distinguished from improvements attributable to the ontology-guided representation itself. This separation is essential because the AAN corpus is strongly skewed toward the Discuss category, in which the majority class comprises 69.8% of instances and the three smallest classes together account for 5.0% of the data.

We report five conditions, all under the protocol of Section 3.10:

(A) **SciBERT, no balancing**: fine-tuned SciBERT with standard cross-entropy loss and no class-level adjustment.
(B) **SciBERT, class-weighted loss**: SciBERT fine-tuned with class weights set inversely proportional to class frequencies in the training folds.
(C) **SciBERT, SMOTE**: SciBERT fine-tuned with SMOTE applied to the `[CLS]` embeddings of the training folds, before the classification head.
(D) **Proposed, no balancing**: the full four-stream framework with standard cross-entropy loss and no class-level adjustment.
(E) **Proposed, matched balancing**: the full framework with the same balancing strategy applied to its feature vector, run once with class-weighted loss (E1) and once with training-fold SMOTE (E2), matching conditions B and C.

For SMOTE in the proposed framework, oversampling is applied to the concatenated vector v within each training fold. As noted in Section 3.8.1, this produces fractional values in the ontological block $\phi(T)$, so the class-weighted condition is the primary basis for comparison and the SMOTE condition is reported alongside it. Table 8 reports the six conditions.

Under the matched comparison with no balancing (A vs D), the proposed framework improves on SciBERT by 1.7 points, the non-significant parity result reported in Section 4.3.

Under the matched comparison with class-weighted loss (B vs E1), the proposed framework reaches 0.571 against 0.562 for SciBERT, a difference of 0.9 points. This is the primary comparison for the proposed architecture, because class weighting adjusts the loss without altering the ontological feature distribution. The gap is narrower than under no balancing, since both systems gain minority-class recall from the weighting; the proposed representation retains a small margin, which indicates that the multi-stream representation carries information the base classifier does not recover from imbalance adjustment alone.

Table 8 Performance under matched imbalance conditions. Macro-F1 is the primary metric; weighted F1 is shown for comparison. The controlled comparisons are the matched pairs A vs D, B vs E1, and C vs E2.

Condition	Balancing	Macro-F1	Weighted F1
SciBERT (A)	None	0.531	0.724
SciBERT (B)	Class-weighted	0.562	0.731
SciBERT (C)	SMOTE	0.574	0.735
Proposed (D)	None	0.548	0.733
Proposed (E1)	Class-weighted	0.571	0.738
Proposed (E2)	SMOTE	0.618	0.765

Under the matched comparison with SMOTE (C vs E2), the proposed framework reaches 0.618 against 0.574 for SciBERT, a difference of 4.4 points. This comparison is not exactly like-for-like: SMOTE is applied to the SciBERT `[CLS]` embeddings in condition C and to the concatenated vector v in condition E2, and v includes the eight-dimensional ontological block. Part of any difference may reflect this feature-space difference rather than the architecture, so the SMOTE comparison is reported alongside the class-weighted comparison rather than in place of it.

The class-weighted SciBERT baseline (B, 0.562) outperforms the unweighted SciBERT (A, 0.531) by 3.1 points, so imbalance handling is a substantial factor on this corpus. Systems reporting results on the AAN corpus should therefore either apply imbalance mitigation or state its absence.

4.4.1 Per-class performance.

Figure 2 gives per-class precision, recall, and F1 under training-fold SMOTE, derived from the confusion matrix of Figure 4 so that the two are consistent by construction.

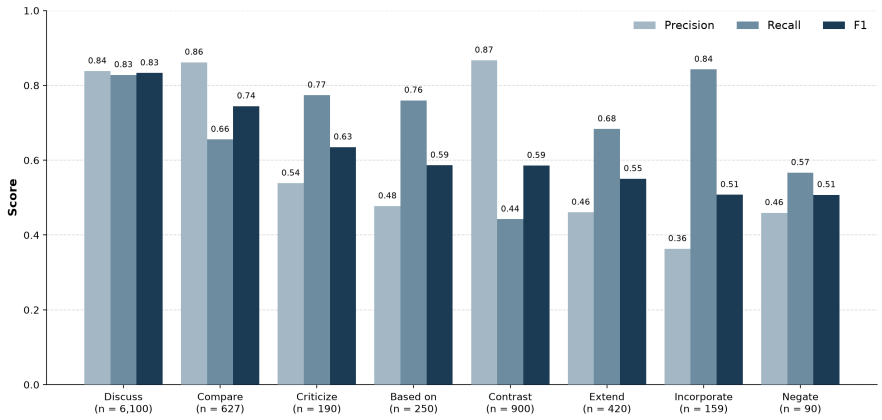

Fig. 2 Per-class precision, recall, and F1 on the eight-class CCRO taxonomy under training-fold SMOTE, ordered by F1. Class support is given beneath each label. Values are computed from the confusion matrix of Figure 4.

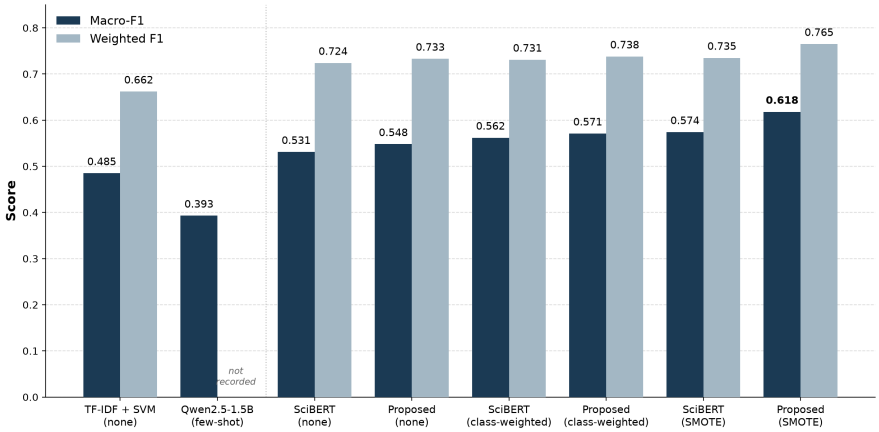

Fig. 3 Macro-F1 and weighted F1 across comparison systems and the proposed framework on the eight-class CCRO taxonomy, averaged over five cross-validation folds and five random seeds. The dotted rule separates the unmatched baselines from the three matched SciBERT/proposed pairs. Values are those of Tables 6, 7 and 8; weighted F1 was not measured for the few-shot baseline.

F1 alone understates what separates the categories, because precision and recall diverge sharply. Discuss is the only category balanced on both (0.838 and 0.828). Contrast pairs the highest precision of the eight (0.867) with the lowest recall (0.442): when the model predicts Contrast it is almost always right, but it recovers under half of the Contrast instances. Incorporate is the mirror image, with the highest recall (0.843) and the lowest precision (0.363). Extend (0.461 precision, 0.683 recall) and Based on (0.477, 0.760) follow the same pattern in weaker form.

This asymmetry is what oversampling produces. SMOTE makes the classifier readier to predict the minority categories, which raises their recall and depresses their precision as majority-class instances are drawn into their columns (Section 4.7). Categories whose F1 is similar can therefore fail for opposite reasons: Contrast and Based on both reach $F1 = 0.586$, the first by missing instances it should recover and the second by claiming instances that belong to Discuss. Figure 3 summarises macro-F1 and weighted F1 across all comparison systems.

These figures should be read with their supports in view (Table 4). Negate has 90 instances and Incorporate 159, so each fold contributes about 18 and 32 respectively, and at these sizes a few-point difference between two minority classes does not establish an ordering.

4.4.2 Stability across seeds.

Table 9 reports the mean and standard deviation of macro-F1 over the five seed-level means for the key configurations, following Section 3.10.2.

The standard deviations quantify the variation contributed by SMOTE synthesis and by the stochastic components of the ensemble. They range from 0.004 to 0.008 and are largest for the SMOTE conditions, consistent with the random choice of synthetic neighbours being the main source of run-to-run variation. The 1.7-point

Table 9 Multi-seed results for the key configurations (mean macro-F1 ± standard deviation over five seeds).

Configuration	Balancing	Macro-F1
SciBERT (baseline)	None	0.531 ± 0.004
SciBERT	Class-weighted	0.562 ± 0.005
SciBERT	SMOTE	0.574 ± 0.006
Proposed	None	0.548 ± 0.006
Proposed	Class-weighted	0.571 ± 0.005
Proposed	SMOTE	0.618 ± 0.008

Table 10 Ablation results (macro-F1, %) on the All-POS feature subset. Rows isolate the drop when a single stream is removed: Stream 1, POS sequences; Stream 2, statistical (TF-IDF) features; Stream 3, contextual embeddings; Stream 4, ontology-derived features.

Configuration	SVM	XGBoost	Ensemble
Full feature set	58.70	57.40	61.80
− Stream 4	56.17	55.17	58.25
− Stream 3	57.20	56.10	59.70
− Stream 2	57.60	56.50	60.40
− Stream 1	58.10	56.90	61.00

A vs D difference is of the same order as this spread, which is consistent with its non-significant p-value; the 4.4-point C vs E2 difference is several times larger than it.

4.5 Ablation analysis

To quantify each stream's contribution, we ablate one stream at a time while retaining the other three, reporting macro-F1 (%) for the SVM, XGBoost, and voting ensemble on the All-POS subset. All rows use training-fold SMOTE, the condition under which the full-feature ensemble reaches 61.80 (condition E2 of Table 8), so each row is read against the full-feature row of its column. Because SMOTE introduces fractional values into the ontological block (Section 3.8.1), the ablation measures each stream's marginal contribution within that feature space; what it establishes is the ordering of the streams rather than the absolute size of any margin. Table 10 gives the results.

The ablation results provide evidence that the ontology-derived representation contributes complementary information to the other feature streams. Removing Stream 4 reduces ensemble macro-F1 from 61.80% to 58.25%, the largest single-stream decline among the four representations, and the same ordering holds for the SVM (-2.53 points) and for XGBoost. The margin is reported as an effect size rather than a significance test: the uncorrected statistic for the SVM comparison gives $p = 0.011$, which the variance correction of Section 3.10.3 does not carry through at this fold ratio,

so the claim rests on the ordering being consistent across all three classifiers and on the standalone and pairwise results of Section 4.3.1. Removing the contextual SciBERT stream produces the second-largest reduction, while removal of the POS and TF-IDF streams results in smaller changes. These results suggest that the contextual and ontology-derived representations make the largest contributions, although the ablation design measures marginal rather than uniquely attributable information: leave-one-out ablation quantifies each stream's contribution given the other three and does not separate unique from redundant information.

The per-stream ordering is consistent with the per-class pattern. Removing Stream 3 affects minority classes such as Criticize and Negate most, where contextual representations carry the main discriminative signal. Removing Stream 1 has the smallest overall effect but reduces performance on the positive-intent categories (Extend, Based on), where VerbNet verb-class patterns supply class-specific cues. Read together with the standalone result of Section 4.3.1, in which the trigger matcher alone reaches 0.310, the ablation shows that the ontological stream is weak as a classifier but the largest marginal contributor within the combined representation.

4.6 Instance-level interpretability

This section evaluates the audit trail defined in Section 3.9 on a sample of test instances, so that interpretability is examined as a property of the model's decision process rather than inferred from the architecture.

4.6.1 Sampling protocol.

From the held-out test instances across the cross-validation folds, using the model with class-weighted loss so that the feature space is unaltered, we sample 100 instances at random, stratified by class to preserve the distribution of the eight categories. For each sampled instance we record the gold CCRO class; the predicted class from the full ensemble; the activated CCRO categories and the specific lexical triggers that fired; the three highest-weighted POS-specific features (VerbNet classes and WordNet synsets) from Stream 1; the five highest-weighted TF-IDF terms from Stream 2; and the predicted class and confidence obtained when Stream 4 is removed and, as a comparison, when Stream 3 is removed. From this record we compute interpretability consistency and stream sensitivity as defined in Section 3.9. Table 11 reports both quantities.

Interpretability consistency records how often the final prediction follows the ontology-derived cue when one is present. It is an operational test of whether the weak-supervision signal is used or overridden, and a measure of faithfulness: a reader inspecting the trigger lexicon and observing that the matcher activated *Contrast* can expect the prediction to be *Contrast* unless the contextual or statistical streams supply overriding evidence. Stream sensitivity records the change in prediction confidence when a stream is removed, together with the proportion of instances whose predicted class changes; instances whose class changes when Stream 4 is removed are those on which the ontological evidence is decisive rather than confirmatory.

Table 11 Interpretability consistency and stream sensitivity on 100 stratified test instances. Consistency is computed only over instances with at least one ontological trigger activated.

Metric	Value
Interpretability consistency (overall)	0.72
Consistency, positive classes	0.68
Consistency, negative classes	0.58
Consistency, neutral classes	0.78
Mean confidence, all streams	0.81 ± 0.11
Mean confidence, Stream 4 removed	0.76 ± 0.13
Mean confidence, Stream 3 removed	0.71 ± 0.14
Predictions changed, Stream 4 removed	22%
Predictions changed, Stream 3 removed	31%

Overall consistency is 0.72: in roughly three-quarters of the instances carrying an ontological signal, the prediction follows the ontology-derived cue. Consistency is lowest on the negative classes (0.58) and highest on the neutral classes (0.78). The negative figure tracks the per-class results of Section 4.4.1. Negate has both the sparsest trigger coverage in the dictionary (12 entries) and the smallest support (90 instances), so on that class the classifier relies more on contextual evidence and diverges more often from the ontological cue.

The consistency figures are computed on a class-stratified sample, not a naturally distributed one. On the corpus distribution, where Discuss is 69.8% of instances, the rate at which the ontological stream changes an outcome would be lower. The per-polarity figures rest on few negative-class instances and are indicative rather than precise.

Mean confidence falls from 0.81 to 0.76 when Stream 4 is removed ($\Delta_4 = 0.05$) and to 0.71 when Stream 3 is removed ($\Delta_3 = 0.10$), so the contextual representation remains the larger driver of the decision. Removing Stream 4 nonetheless changes the predicted class on 22% of instances, against 31% for Stream 3. The ontological evidence is therefore not only a confidence modifier: on roughly a fifth of instances it determines which class is predicted.

4.6.2 Diagnostic audit trail.

The same record can be inspected for an individual prediction. For each instance it holds the activated CCRO categories and the triggers that fired, the Stream 1 and Stream 2 features carrying the most weight, the predicted class and its confidence, and the class and confidence the model assigns when Stream 4 or Stream 3 is withheld. Reading those fields together identifies which representation supplied the evidence for a given decision, and whether the ontological prior reinforced that decision, determined it, or was silent. Table 12 gives the full record for three AAN instances on which the model errs, where the diagnostic value of the trail is clearest.

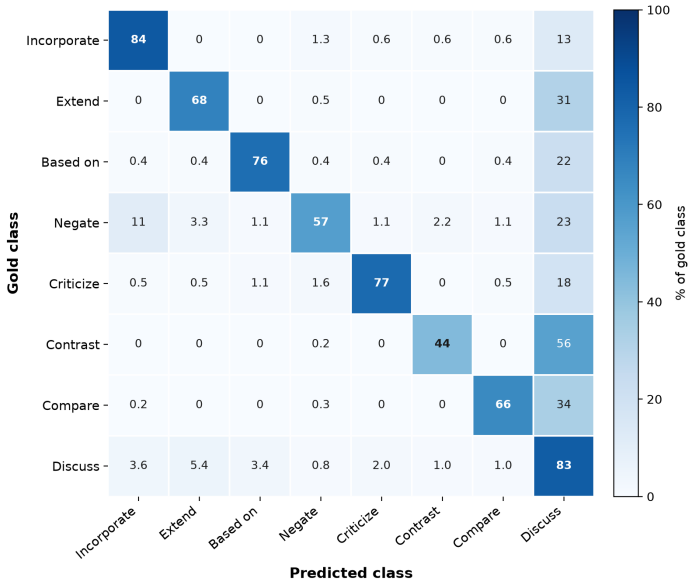

Fig. 4 Normalised confusion matrix on the eight-class CCRO taxonomy under training-fold SMOTE (condition E2 of Table 8). Each row gives the percentage of instances of that gold class assigned to each predicted class; rows sum to 100. Diagonal entries are per-class recall.

The consistency metric excludes instances on which no trigger fires. For those instances interpretability rests on Stream 1 or Stream 2 rather than on the ontology, and Section 4.7 examines the citation contexts in which the trigger dictionary has no coverage.

4.7 Error analysis

Two mechanisms account for most errors. The first is loss to the majority category: under high syntactic ambiguity the classifier assigns Discuss, which holds 69.8% of the corpus (Table 4) and is the most probable assignment without a discriminating signal. The second is realisation through noun phrases (e.g. *the incorporation of*) rather than the verb-phrase patterns that activate VerbNet mappings, which degrades Stream 1 and leaves the trigger matcher silent; W08-2102 in Table 12 is an instance. The two act together on the smallest categories, where lexical overlap with Discuss and data scarcity compound.

Figure 4 gives the confusion matrix for the full model under training-fold SMOTE. The error mass is concentrated almost entirely in exchanges with Discuss, in both directions. Confusion among the seven remaining categories is close to absent: setting aside the Discuss row, only Negate sends more than 2% of its instances to a category other than Discuss, and that cell (11.1% to Incorporate) is ten instances.

In the gold-to-Discuss direction, Contrast is the most affected category. It loses 55.6% of its instances (500 of 900) to Discuss and reaches a recall of 0.442, the lowest of the eight; W07-0716 in Table 12 is an instance of this cell. Compare loses 34.0%

Table 12 Error analysis: representative misclassifications with the stream-level audit trail. Each prediction cell gives the predicted class and the soft-voting confidence. Stream 4 is the ontology-derived block, Stream 3 the contextual embeddings; each ablated model is retrained on the same training folds.

Reference ID	Citation context (source text)	Gold	Full model	Without Stream 4	Without Stream 3
W08-2102 (3)	*"The method shows improvements over the method described in,[45] . . . similar to that of,[46] suggesting that the incorporation of constituent structure can improve dependency accuracy."*	Incorporate	Discuss (0.68)	Discuss (0.64)	Discuss (0.52)
W07-0716 (10)	*"Och showed that system performance is best when parameters are optimised using the same objective function that will be used for evaluation; this remains common . . . even when alternative evaluation measures are used.[47]"*	Contrast	Discuss (0.71)	Discuss (0.67)	Discuss (0.58)
W08-0320 (11)	*"We used these weights in a beam search decoder to produce translations for the test sentences, which we compared to the WMT07 gold standard.[47]"*	Based on	Compare (0.74)	Compare (0.70)	Discuss (0.61)

to Discuss, Extend 31.2%, Negate 23.3%, Based on 22.0%, and Criticize 18.4%. The ordering matches the account above: where the discriminating signal is weak, the classifier assigns the majority category.

The reverse direction is where oversampling shows. Discuss sends only small percentages to the minority categories, but because it holds 6,100 instances those percentages are large counts: 331 instances to Extend, 222 to Incorporate, 205 to Based on, and 123 to Criticize. They dominate the minority columns. Of the 369 instances predicted Incorporate, 222 (60%) are in fact Discuss, and of the 623 predicted Extend, 331 (53%) are Discuss. This is why Incorporate combines a recall of 0.843 with a precision of 0.363. Training-fold SMOTE raises minority recall by making the classifier readier to predict minority categories and pays for it in minority precision; the per-class F1 values of Section 4.4.1 are the net of the two effects, which is why a category can hold high recall and still report a modest F1.

The matrix supports the qualitative account while sharpening it. The dominant failure is not diffuse confusion among semantically adjacent intents, which the near-empty off-diagonal blocks rule out, but a specific exchange with the majority category that runs in both directions once oversampling is applied.

Because each stream is separable, a misclassification can be attributed to a specific mechanism. Table 12 gives the per-instance audit trail defined in Section 3.9 for three representative AAN failures, recording the prediction and confidence of the full model alongside those obtained with Stream 4 and Stream 3 withheld, and the stream whose evidence was absent or misleading.

The failure mechanisms are as follows. For W08-2102 (3), Incorporate overlaps lexically with Discuss and shares verb-noun clusters with it; the intent is realised by the noun phrase *the incorporation of* rather than by a verb, so no VerbNet verb-class mapping is available for Stream 1 to activate on and no trigger fires in Stream 4. For W07-0716

(10), clause-long descriptions and non-standard linguistic structures cause boundary resolution to fail, and where ambiguity is high the system reverts to the majority Discuss class. For W08-0320 (11), the surface comparative marker *compared to* is lexically salient, while the methodological dependency that determines the gold label is expressed structurally rather than by any trigger term, so the prediction follows the surface cue.

The trail reads consistently across the three. Withholding Stream 4 lowers confidence by 0.04 in every case but changes no predicted class, so none of these errors is caused by the ontological prior and none is repaired by it. Withholding Stream 3 costs between 0.13 and 0.16, which places these instances on the same side of the aggregate pattern of Section 4.6, where Δ_3 is roughly twice Δ_4.

The third instance is the most informative. If the trigger dictionary were responsible for the Compare prediction, removing Stream 4 would release it; the prediction instead stays Compare at 0.70. The comparative cue driving the error is therefore carried by the statistical and contextual streams rather than by the ontology, and withholding Stream 3 collapses the prediction to the majority class. This is the kind of attribution the modular design is for: the stream responsible for an error is identified by removal rather than inferred from an approximation of the model.

5 Discussion

5.1 Predictive performance

The central finding of this study is that explicit citation semantics can be incorporated into a contextual classification pipeline without sacrificing predictive performance relative to a fine-tuned scientific language model. The proposed representation does not establish a statistically significant accuracy advantage over SciBERT under the matched condition. Its value instead emerges from the combination of predictive parity, stream-level analysis, and explicit semantic evidence that can be inspected during classification. This distinction is important because citation-intent classification is not only a prediction problem: its outputs may subsequently be used to construct semantic scholarly networks, support literature discovery, or analyse relationships among scientific contributions. In such settings, understanding the basis of a classification can be as important as obtaining the classification itself.

The matched-balancing experiment (Section 4.4) supports a second reading. Class weighting improves the SciBERT baseline substantially, which indicates that a share of the differences reported for this corpus in earlier work is attributable to imbalance treatment rather than to representation. Separating the two is a precondition for interpreting any architectural claim on the AAN corpus.

5.2 Contribution of the ontology-derived representation

Three results bear on the role of the ontological stream. The standalone trigger matcher reaches a macro-F1 of 0.310, so the lexical rules are not discriminative enough to classify citation intent by themselves. Pairing the same indicators with SciBERT embeddings raises macro-F1 from 0.531 to 0.545, so the indicators carry signal the

contextual representation does not already encode. Removing the stream from the full model produces the largest single-stream decline of the four, from 61.80% to 58.25% for the ensemble. The instance-level analysis gives the same result at the level of individual decisions: removing the stream changes the predicted class on 22% of sampled instances. Together these establish that the classifier draws on the ontological stream rather than defaulting to the much larger embedding block, which was the empirical question raised in Section 3.6.

The scope of that claim should be stated precisely. The CCRO trigger dictionary and the annotation scheme derive from the same conceptual taxonomy, so the stream functions as an explicit prior of the weak-supervision kind rather than as an independent source of evidence about the gold labels. What the results establish is that an explicit, inspectable prior contributes measurably within a contextual pipeline, not that ontological knowledge is independently predictive. Ablation also measures marginal contribution given the remaining streams and cannot separate unique from redundant information.

Per-class results do not follow trigger-lexicon coverage in any simple way. The two weakest categories by F1, Negate (0.507) and Incorporate (0.508), are also the two smallest and among the most sparsely triggered, so coverage and training-set size are confounded and neither can be isolated without holding the other fixed. Compare, however, has few trigger entries and the second-highest F1 of the eight (0.745), so sparse coverage is not on its own a predictor of weak per-class performance. We therefore treat the trigger counts as an observation about where the ontological stream has least to contribute, not as an explanation of the per-class ordering.

5.3 Instance-level interpretability

The ontology-guided representation provides an explicit semantic pathway through which citation-intent predictions can be examined. For each evaluated citation context, the system records the activated ontology category, matched lexical evidence, and the prediction obtained when the ontology stream is removed. This allows interpretability to be examined as a property of the model's decision process rather than inferred solely from its architecture. Cases in which removal of the ontology stream changes the predicted class or substantially reduces prediction confidence provide direct evidence that the explicit semantic representation is being used by the classifier.

The analysis also reveals cases in which ontology-derived evidence is insufficient. Citation contexts expressing methodological relationships through noun phrases or long syntactic constructions may fail to activate the relevant lexical triggers, as in the Incorporate and Compare failures of Section 4.7. Such cases demonstrate both the usefulness and the limits of ontology-guided representation.

5.3.1 Faithful attribution through model components.

The proposed approach differs from post-hoc explanation methods in that the semantic evidence used for classification is represented explicitly as part of the model input. This permits attribution to be examined by controlled removal of individual representation streams. Such analysis does not imply that architectural interpretability is

universally superior to post-hoc explanation; rather, it provides a direct means of testing whether the explicit semantic representation affects model behaviour. The distinction is particularly relevant here because the study's objective is to expose the role of citation-semantic knowledge in classification rather than to provide a generic explanation for an otherwise opaque model.

The property claimed is faithfulness. Interpretability names several conflicting properties,[48] and evaluation must state which is claimed and against what criterion.[49] A post-hoc method fits a second model that approximates the first, with no guarantee that the features it marks as important are the features the computation used; a plausible explanation may be unfaithful.[50] Removing a physically separate feature block and measuring the effect instead operates on the model itself, which is the argument for models interpretable by construction where a decision carries consequences,[51] as research evaluation does.

5.4 Implications for scholarly information analysis

The framework supports several potential scholarly-information applications. First, classified citation relationships can be represented as typed edges in scholarly knowledge graphs, allowing relationships such as methodological extension, comparison, and criticism to be distinguished rather than represented as undifferentiated citation links. Second, citation intent can support semantic literature discovery by enabling users to identify papers that, for example, extend a method, provide a methodological basis, or critically evaluate prior work. Third, citation-intent information can complement conventional citation indicators in research-analysis systems by distinguishing different forms of scholarly influence. These applications are presented as potential uses of the generated semantic information rather than as validated downstream tasks in the present study.

5.5 Ethical considerations

Automated citation-intent classification carries bias and representation risks.[23] The AAN corpus is severely imbalanced (Discuss 69.8%, neutral 77.0%), reflecting computational-linguistics practice that may not transfer elsewhere, and historical data risks reproducing scientometric biases: citation practice disadvantages non-English research,[52] English dominance lowers the visibility of non-anglophone and Global South work,[53] and gendered citation imbalances are documented.[54] Since the classifier learns these patterns, indicators used for funding or tenure could propagate them; equity-conscious modelling, transparent limitation reporting, and broader annotation are needed for fair deployment.

5.6 Limitations

Several limitations should be considered when interpreting the findings. First, the eight-class CCRO taxonomy imposes a single-label representation on citation contexts that may express multiple or overlapping intentions. Second, the annotations were produced by graduate researchers and therefore represent a consensus-based gold standard rather than an independently adjudicated expert reference.[40] Third, errors

in upstream linguistic processing may propagate to the POS- and ontology-derived representations; spaCy POS-tagging was validated at approximately 89.7% on this corpus and errs mainly on technical terms. Fourth, the ontology-derived lexical representation is based on a predefined dictionary whose conceptual basis is related to the annotation taxonomy. This is a property of the weak-supervision design rather than a defect, but it means the stream's contribution should not be read as independent evidence of the gold labels. Fifth, the current fusion mechanism uses feature concatenation rather than learned ontology-guided interaction.[55, 56] Finally, evaluation is confined to the AAN corpus, which is drawn primarily from computational linguistics and is strongly imbalanced; multi-domain and cross-taxonomy testing remains future work. The annotated CCRO labels are available from the corresponding author on request rather than deposited publicly, which constrains independent replication.

6 Conclusion

This study investigated ontology-guided citation intent classification as a means of combining contextual language representations with explicit scholarly citation semantics. The proposed framework integrates linguistic, statistical, neural, and CCRO-derived representations and was evaluated on 8,736 manually annotated citation contexts using an eight-class citation-intent taxonomy. Under matched conditions, the framework achieved predictive performance comparable to a fine-tuned SciBERT baseline, while ablation analysis demonstrated that the ontology-derived representation made the largest marginal contribution among the individual feature streams.

The principal contribution is therefore not a claim of superior classification accuracy, but an approach for incorporating explicit citation semantics into a computational pipeline in a form that can be examined through controlled feature-stream analysis. This provides a basis for more transparent citation-intent classification and for downstream scholarly-information applications such as semantic literature discovery and citation-aware knowledge graphs.

The findings also identify clear directions for further development, including learned ontology-guided fusion, broader multi-domain evaluation, improved treatment of overlapping citation intents, and richer instance-level evaluation of explanation faithfulness. Overall, the results support the feasibility of combining structured citation knowledge with contextual scientific language representations for interpretable scholarly information analysis.

Acknowledgements. Not applicable.

Declarations

Funding. The authors received no financial support from any public, commercial, or not-for-profit funding agency for the research, authorship, and/or publication of this article. The research was carried out as an independent academic exercise by researchers from the National University of Modern Languages, the Institute of Space Technology, and Harbin Engineering University.

Conflict of interest. The authors declare no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.
Ethics approval and consent to participate. Not applicable.
Data availability. The data produced and processed in the present study, a collection of 8,736 sentences from research papers in the ACL Anthology Network (AAN) corpus labelled with citation intent, are not publicly shared, in line with the data-sharing regulations of the National University of Modern Languages. The data can be requested from the corresponding author on reasonable demand and with formal approval from the universityś research administration office.
Author contribution. First Author: formal analysis, software, implementation. Second Author: conceptualization, methodology, writing — original draft. Third Author: manuscript outlining, writing — review and editing, proof-reading.
Use of generative AI. During the preparation of this manuscript the authors used AI assistance to improve readability and language. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Mustafa, G., Rauf, A., Ahmed, B., Afzal, M., Akhunzada, A., Alharthi, S.: Comprehensive evaluation of publication and citation metrics for quantifying scholarly influence. IEEE Access **11** (2023) https://doi.org/10.1109/ACCESS.2023.3290917

[2] Roman, M., Shahid, A., Khan, S., Koubaa, A., Yu, L.: Citation intent classification using word embedding. IEEE Access **9** (2021) https://doi.org/10.1109/ACCESS.2021.3050547

[3] Meng, R., Lu, W., Chi, Y.-h., Han, S.: Automatic Classification of Citation Function by New Linguistic Features. Semantic Scholar CorpusID:196103130 (2017). https://api.semanticscholar.org/CorpusID:196103130

[4] Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. 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), pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1371

[5] Cohan, A., Ammar, W., Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 3586–3596. Association for Computational Linguistics, Minneapolis, MN (2019). https://doi.org/10.18653/v1/N19-1361

[6] Fogelson, A., Trišović, A., Thompson, N.: LLMs in citation intent classification:

Progress, precision, and reproducibility challenges. In: Proceedings of the 3rd ACM Conference on Reproducibility and Replicability (ACM REP '25) (2025). https://doi.org/10.1145/3736731.3746137

[7] Paolini, L., Vahdati, S., Di Iorio, A., Wardenga, R., Heibi, I., Peroni, S.: CiteFusion: An ensemble framework for citation intent classification harnessing dual-model binary couples and SHAP analyses. Scientometrics **130**(11), 5911–5981 (2025) https://doi.org/10.1007/s11192-025-05418-8

[8] Karaca, K.S., Eravci, B.: A large-scale dataset and citation intent classification in turkish with LLMs. In: 2025 10th International Conference on Computer Science and Engineering (UBMK), pp. 509–514. IEEE, Piscataway, NJ, USA (2025). https://doi.org/10.1109/UBMK67458.2025.11207038

[9] Touza, I., Balama, G., Warda, L., Kaladzavi, G., Kolyang: Ontology-driven text classification and data mining: Beyond keywords toward semantic intelligence. Revue d'Intelligence Artificielle **39** (2025) https://doi.org/10.18280/ria.390301

[10] Ihsan, I., Qadir, M.A.: CCRO: Citation's context and reasons ontology. IEEE Access **7** (2019) https://doi.org/10.1109/ACCESS.2019.2903450

[11] Moravcsik, M.J., Murugesan, P.: Some results on the function and quality of citations. Social Studies of Science **5**(1), 86–92 (1975) https://doi.org/10.1177/030631277500500106

[12] Chubin, D.E., Moitra, S.D.: Content analysis of references: Adjunct or alternative to citation counting? Social Studies of Science **5**(4), 423–441 (1975) https://doi.org/10.1177/030631277500500403

[13] Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI), pp. 926–931 (1999)

[14] Garzone, M., Mercer, R.E.: Towards an automated citation classifier. In: Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, pp. 337–346. Springer, Berlin, Heidelberg (2000). https://doi.org/10.1007/3-540-45486-1_28

[15] Nakov, P.I., Schwartz, A.S., Hearst, M.A.: Citances: Citation sentences for semantic analysis of bioscience text. In: Proceedings of the SIGIR'04 Workshop on Search and Discovery in Bioinformatics, pp. 81–88 (2004)

[16] Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 103–110. Association for Computational Linguistics, Sydney, Australia (2006)

[17] Abu-Jbara, A., Ezra, J., Radev, D.: Purpose and polarity of citation: Towards NLP-based bibliometrics. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 596–606. Association for Computational Linguistics, Atlanta, GA (2013)

[18] Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D.: Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics **6**, 391–406 (2018) https://doi.org/10.1162/tacl_a_00028

[19] Hwang, M., Choi, C., Kim, P.: Automatic enrichment of semantic relation network and its application to word sense disambiguation. IEEE Transactions on Knowledge and Data Engineering **23**(6), 845–858 (2011)

[20] Liu, M.-C.: The construction and annotation of a semantically enriched database: The mandarin VerbNet and its NLP applications. In: Chinese Lexical Semantics, pp. 257–272. Springer, Singapore (2020). https://doi.org/10.1007/978-981-32-9240-6_18

[21] Wu, H.C., Luk, R.W.P., Wong, K.-F., Kwok, K.-L.: Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems **26**(3), 13–11337 (2008)

[22] Peroni, S., Shotton, D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics **17**, 33–43 (2012) https://doi.org/10.1016/j.websem.2012.08.001

[23] Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Cohan, A., Jurgens, D., Lo, K.: MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 1875–1889. Association for Computational Linguistics, Seattle, WA (2022). https://doi.org/10.18653/v1/2022.naacl-main.137

[24] Zhang, Y., Wang, Y., Sheng, Q.Z., Yao, L.: Multi-task learning model for citation intent classification in scientific publications. Scientometrics **128**(12) (2023) https://doi.org/10.1007/s11192-023-04858-4

[25] Buscaldi, D., Dessì, D., Motta, E., Osborne, F., Reforgiato Recupero, D.: Mining scholarly publications for scientific knowledge graph construction. In: The Semantic Web: ESWC 2019 Satellite Events, pp. 8–12. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32327-1_2

[26] Kunnath, S.N., Herrmannova, D., Pride, D., Knoth, P.: A meta-analysis of semantic classification of citations. Quantitative Science Studies **2**(4), 1170–1215 (2021)

https://doi.org/10.1162/qss_a_00159

[27] Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)

[28] Schuler, K.K.: VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania (2005)

[29] Miller, G.A.: WordNet: A lexical database for english. Communications of the ACM **38**(11), 39–41 (1995) https://doi.org/10.1145/219717.219748

[30] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[31] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[32] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[33] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002) https://doi.org/10.1613/jair.953

[34] Blagus, R., Lusa, L.: SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics **14**, 106 (2013) https://doi.org/10.1186/1471-2105-14-106

[35] Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation **10**(7), 1895–1923 (1998) https://doi.org/10.1162/089976698300017197

[36] Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning **52**(3), 239–281 (2003) https://doi.org/10.1023/A:1024068626366

[37] Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979)

[38] Dror, R., Baumer, G., Shlomov, S., Reichart, R.: The hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1383–1392 (2018). https://doi.org/10.18653/v1/P18-1128

[39] Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. Language Resources and Evaluation **47**(4), 919–944 (2013) https://doi.org/10.1007/s10579-012-9211-2

[40] Pride, D., Knoth, P.: An authoritative approach to citation classification. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 337–340 (2020). https://doi.org/10.1145/3383583.3398617

[41] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971) https://doi.org/10.1037/h0031619

[42] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960) https://doi.org/10.1177/001316446002000104

[43] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977) https://doi.org/10.2307/2529310

[44] Yang, A., Yang, B., Zhang, B., et al.: Qwen2.5 Technical Report (2024). https://doi.org/10.48550/arXiv.2412.15115

[45] Koo, T., Carreras, X., Collins, M.: Simple semi-supervised dependency parsing. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 595–603. Association for Computational Linguistics, Columbus, OH (2008)

[46] McDonald, R., Pereira, F.: Online learning of approximate dependency parsing algorithms. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 81–88. Association for Computational Linguistics, Trento, Italy (2006)

[47] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, PA (2002). https://doi.org/10.3115/1073083.1073135

[48] Lipton, Z.C.: The mythos of model interpretability. Communications of the ACM **61**(10), 36–43 (2018) https://doi.org/10.1145/3233231

[49] Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). https://doi.org/10.48550/arXiv.1702.08608

[50] Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4198–4205 (2020). https://doi.org/10.18653/v1/2020.acl-main.386

[51] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019) https://doi.org/10.1038/s42256-019-0048-x

[52] Amano, T., González-Varo, J.P., Sutherland, W.J.: Languages are still a major barrier to global science. PLoS Biology **14**(12), 2000933 (2016) https://doi.org/10.1371/journal.pbio.2000933

[53] Meneghini, R., Packer, A.L.: Is there science beyond english? initiatives to increase the quality and visibility of non-english publications might help to break down language barriers in scientific communication. EMBO Reports **8**(2), 112–116 (2007) https://doi.org/10.1038/sj.embor.7400906

[54] Dworkin, J.D., Linn, K.A., Teich, E.G., Zurn, P., Shinohara, R.T., Bassett, D.S.: The extent and drivers of gender imbalance in neuroscience reference lists. Nature Neuroscience **23**(8), 918–926 (2020) https://doi.org/10.1038/s41593-020-0658-y

[55] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019) https://doi.org/10.1109/TPAMI.2018.2798607

[56] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML), pp. 689–696 (2011)