
CiteFusion: An Ensemble Framework for Citation Intent Classification Harnessing Dual-Model Binary Couples and SHAP Analyses

Lorenzo Paolini (ORCID: 0009-0003-3803-4011)

Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Sahar Vahdati (ORCID: 0000-0002-7171-169X)

Nature-inspired machine intelligence group, SCaDS.AI center, Technical University of Dresden, Germany

Institute for Applied Computer Science, InfAI - Dresden, Germany

Angelo Di Iorio (ORCID: 0000-0002-6893-7452)

Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

Robert Wardenga (ORCID: 0009-0004-3317-6122)

Institute for Applied Computer Science, InfAI - Dresden, Germany

Ivan Heibi (ORCID: 0000-0001-5366-5194)

Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Silvio Peroni (ORCID: 0000-0003-0530-4305)

Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Digital Humanities Advanced Research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Contacts: lorenzo.paolini11@unibo.it

Abstract

Understanding the motivations underlying scholarly citations is critical for evaluating research impact and fostering transparent scholarly communication. This study introduces *CiteFusion*, an ensemble framework designed to address the multiclass Citation Intent Classification (CIC) task on benchmark datasets, SciCite and ACL-ARC. The framework decomposes the task into binary classification subtasks, utilizing complementary pairs of SciBERT and XLNet models fine-tuned independently for each citation intent. These base models are aggregated through a feedforward neural network meta-classifier, ensuring robust performance in imbalanced and data-scarce scenarios. To enhance interpretability, SHAP (SHapley Additive exPlanations) is employed to analyze token-level contributions and interactions among base models, providing transparency into classification dynamics. We further investigate the semantic role of structural context by incorporating section titles into input sentences, demonstrating their significant impact on classification accuracy and model reliability. Experimental results show that *CiteFusion* achieves state-of-the-art performance, with Macro-F1 scores of 89.60% on SciCite and 76.24% on ACL-ARC. The original intents from both datasets are mapped to Citation Typing Ontology (CiTO) object properties to ensure interoperability and reusability. This mapping highlights overlaps between the two datasets labels, enhancing their understandability and reusability. Finally, we release a web-based application that classifies citation intents leveraging *CiteFusion* models developed on SciCite.

Keywords: Citation Intent Classification, Language Models, Ensemble Strategies, Explainable AI

1 Introduction

Assessing research is crucial in scholarly communication as it ensures the quality, relevance, and impact of scientific contributions, fostering an environment of accountability and continuous improvement. By evaluating research, scholars and institutions can identify significant advancements, recognize influential work, and effectively allocate resources to areas with the highest potential for innovation and societal benefit. Moreover, it supports the Open Science movement by promoting transparency, collaboration, and accessibility in research. Specifically, the application of citation analysis and bibliometrics as tools for evaluating research, and consequently for allocating research funding, occupies a central position in this process (Pride, 2022).

However, many researchers have criticized the use of bibliometrics in research assessment, in particular shedding light on the paradoxical nature hidden behind the abuse of metrics involving citation counts as a proxy for research assessment. Pride (2022) states that there are no sufficient evidences to demonstrate a connection between research quality and citation rates, while Wallin (2005) accounts for the widespread use of the *Journal Impact Factor (JIF)* and of the *h-index* - two bibliometrics strongly relying on citation counts - as research quality proxies. Such citation-count-based metrics are used to evaluate researchers, publications, and journals (Li & Ho, 2008), staying unclear on the differentiation behind the types of citations. Indeed, some citations may indicate the reuse of a methodology while some others may merely serve as an acknowledgment of a prior work (Cohan et al., 2019).

Differentiating the nature of citations is thus instrumental in providing more comprehensive and meaningful analyses in research assessment related fields (Small, 2018), and developing tools capable of retrieving influential papers, beyond citation counts, is fundamentally important also to promote more conscious research (Ritchie, 2008). Other possibilities within the field are related to the development of applications for enhanced information retrieval (Moravcsik & Murugesan, 1975; Pride, 2022), document summarization (Cohan & Goharian, 2015), and finally citations occupy a central role also in studies related to the evolution of scientific fields (Jurgens et al., 2018) where they are employed to construct citation networks and to frame different periods.

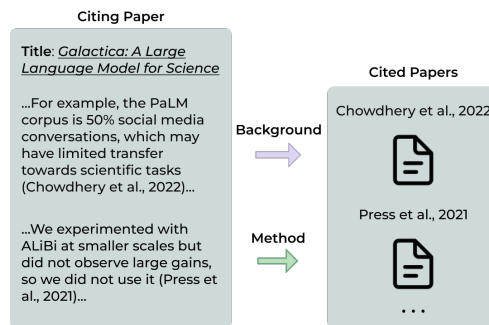


Figure 1. Example of an entity (citing paper) with two different in-text citations, each referring to distinct entities (cited papers) for different intents: Background and Method.

Recognizing the importance to meaningfully distinguish between citation intents to advance scholarly communication also highlights the need for automated methods to accomplish this. The task is known in literature as *Citation Intent Classification (CIC)*, represented in **Figure 1**. Indeed, as highlighted by Pride (2022), “the sheer volume of new research now being produced on an annual basis is far beyond the capacity of a single researcher to investigate even the narrowest of domains without effective search tools”, and this huge amount of new research and consequent citations being produced makes manual classification extremely time consuming and difficult. Instead, automated systems can process large datasets efficiently, providing useful insights into citation behaviors and helping in the discovery of patterns and trends that may be overlooked through manual analyses. In this context, Machine Learning (ML) based solutions become pivotal.

Building on recent progress in ML and Natural Language Processing (NLP), in this work we explore how *Pre-trained Language Models* (PLMs) and *Ensemble Strategies* (ES) can enhance performances on *SciCite* (Cohan et al., 2019) and *ACL-ARC* (Jurgens et al., 2018), two benchmark datasets for CIC characterized by a skewed data distribution. To deal with the imbalance present in these datasets, we reduce the multiclass CIC task to multiple binary tasks, leveraging couples of different PLM architectures to capture both the syntactical and semantical aspects defining each specific citation intent. The outputs of these couples are finally stacked to produce an *Ensemble Classifier* (EC) employing a *Feed Forward Neural Network* (FFNN) head, intended to group back the multiple binary outputs into a coherent multiclass framework, thereby ensuring that the final predictions span the entire range of citation intent categories of each dataset. Despite the use of a FFNN as *main* head, we also investigate different options, both supervised and unsupervised, to see how base PLMs’ predictions can be aggregated. Furthermore, the original labels of SciCite and ACL-ARC are mapped to standardized object properties from the *Citation Typing Ontology* (CiTO), designed by Peroni and Shotton (2012), to enhance reusability.

Furthermore, while prior research highlights structural cues – i.e., section titles – as integral components of multitask learning frameworks (Cohan et al., 2019), we argue that the semantic role of these elements warrants further investigation. Specifically, we investigate whether incorporating the explicit semantic meaning of section titles – beyond their syntactic function – into the classification pipeline can enhance a model’s ability to accurately perceive and categorize citations. For instance, when raw contextual signals – such as individual words or elements within citation contexts – are ambiguous, we argue that a model might rely on the inherent semantic associations of section titles (e.g., "*Methodology*" or "*Results*") and contextual cues to infer citation intents. This approach introduces a novel dimension for improving classification accuracy and interpretability by explicitly encoding the functional role of sections within citations. Thereby, to assess the influence of section titles as semantical components, we develop two ECs for each of the two datasets, resulting in four models: two of them trained with section titles included within the input citation sentences (WS setting), and two of them trained to solve the task through raw citation contexts (WoS setting). Furthermore, the ECs derived from our experiments on the SciCite dataset are integrated into a web-based application designed to classify citation intents.

In this study, we also employ granularly evaluated training loops to overcome the overfitting problem that usually arise when dealing with PLMs fine-tuning (Zheng et al., 2025). Furthermore, since full-parameter fine-tuning with PLMs usually requires substantial time and computational resources, we also employ mixed-precision to reduce them. We perform some computational instability analyses to address both the impact of these techniques and the instability caused by the skewed distributions of SciCite and ACL-ARC datasets on the reproducibility of our FFNN-based ensemble classifiers. Finally, to improve the interpretability and reinforce the reliability of our findings, we incorporate *Explainable AI* (XAI) techniques to give a better understanding of the classification dynamics of the main ECs, and to highlight which features shape the perception of each intent according to different PLM architectures.

Our contributions can be outlined as follows:

- (i) we publicly release four ECs making use of PLMs and FFNNs for citation intent classification, two of them trained on the ACL-ARC dataset and the other two on the SciCite dataset;
- (ii) we surpass the state-of-the-art on the SciCite benchmark;
- (iii) we surpass the state-of-the-art on the ACL-ARC benchmark;
- (iv) we demonstrate the utility of incorporating section titles as semantic components within citation contexts, and discuss how these elements reshape models’ perception of the input sentences;
- (v) we provide a mapping between the original schemas of the two datasets and standardized object properties from CiTO;
- (vi) we release a web-based application for the automatic classification of citation intents, employing the two ECs trained on SciCite.

2 Background and Related Works

Within the field of Citation Intent Classification (CIC), significant advancements and developments have recently been produced, mainly thanks to an increasing spectrum of available methodologies and theoretical frameworks. This section will present the theoretical ground-base to proficiently understand our work. It will start with a discussion on Citation Intent Classification schemas, to then provide an overview on the datasets available for the task. Following this, the section will briefly present historical approaches to the task, together with the main technological innovations that advanced the field. Finally, it will give a bird-eye view on the methodologies employed in this study.

2.1 A feasible Citation Intent Classification Schema

In the field of CIC, it is important to consider the selection of an appropriate classification schema. Specifically, the choice of labels and the number of classes included in such a schema play a critical role in determining both the overall utility and impact of the resulting dataset and the outcomes achievable through automated classification methods. This discourse traces its roots back to the seminal work “*Can Citation Indexing Be Automated?*”, by Garfield (1964), where the author identifies 15 reasons for which a scholar may decide to cite another work, laying the foundations for further research in this domain. Following this foundational contribution many researchers worked around the definition of a meaningful and accurate citation classification scheme (Kunnath et al., 2021). This emphasis on good citation classification schemas is rooted in the diverse functions that citations serve in scholarly communication. These functions include, but are not limited to, acknowledging the source of ideas, evidencing arguments, illustrating methodological similarities, and connecting related academic discussions. A well-designed schema should capture these intricate differences, enabling accurate analyses and interpretation of citation contexts.

The design of a robust schema for annotating citations according to their purpose – or function, or intent – benefits numerous applications. Among them, as stated also in the introduction of this work, citation analysis for research evaluation is a key operationalization (Jochim & Schutze, 2012; Pride, 2022). Nonetheless, the importance of such a schema also extends to numerous areas beyond research evaluation, like academic writing and literature reviews, where understanding the intent behind citations can clarify how each referenced work enriches the academic discourse, providing a more conscious way to engage with it.

Furthermore, developing a description logic schema compatible with semantic web technologies, can unlock the potential to treat bibliographic references, citation contexts, and even rhetorical elements within scientific publications as semantic metadata, which enables better organization, search, and integration for web-based scientific portals (Ciccarese et al., 2014). Building a shared language for scientific discourse through these schemas transforms research articles’ components into a rich web of semantic data, which allows researchers to explore not only by keywords, but by relations. A leading citation classification model is the *Citation Typing Ontology (CiTO)*, an OWL ontology specifically designed to capture both the factual and rhetorical purposes of citations in scholarly works (Peroni & Shotton, 2012). CiTO offers a rich vocabulary of 41 properties, enabling precise annotation of citation semantics.

CiTO also facilitates the alignment of properties between different classification schemas, enabling interoperability and enriched representation of citation intents. This is particularly relevant for the schemas employed in this work. Indeed, both SciCite (Cohan et al., 2019) and ACL-ARC (Jurgens et al., 2018) datasets characterize their citation contexts into two newly designed schemas (SciCite employs three labels, while ACL-ARC utilizes six distinct categories). By mapping these two schemas to CiTO, it becomes possible to identify commonalities and overlaps in their respective intents, thereby fostering a more unified approach to citation intent classification.

2.2 Datasets for Citation Intent Classification

A key consequence of designing an effective citation classification schema is the development of robust and comprehensive datasets. Most of the current datasets designed for classifying citation contexts use their own specific and newly developed schema. While creating multiple specialized schemas drives innovation, it complicates the integration of different resources (Cohan et al., 2019). Moreover, the field still suffers from a shortage of extensive and diverse corpora, mainly because manual annotation of citation contexts according to their intents is labor-intensive, and also requires domain-specific expertise (Pride, 2022). As a result, the limited number of manually annotated datasets currently available cannot yet serve as a definitive gold standard for the CIC task, though they offer important starting points.

Dataset	Categories (distribution)	Source	#papers	#instances
ACL-ARC	Background (0.51)	Computational	186	1,941
	Extends (0.04)	Linguistics		
	Uses (0.19)			
	Motivation (0.05)			
	Compare/Contrast (0.18)			
SciCite	Future work (0.04)		6,627	11,020
	Background (0.58)	Computer		
	Method (0.29)	Science &		
	Result comparison (0.13)	Medicine		

Table 1. Comparison of SciCite and ACL-ARC datasets. Table from (Cohan et al., 2019).

An early contribution to CIC is *ACL-ARC* (Jurgens et al., 2018), which provides nearly 2,000 manually annotated citation contexts from the NLP domain. Despite employing a useful six-labels schema, ACL-ARC’s narrow disciplinary focus and relatively small size limit its generalizability and overall utility. In contrast, the development of *SciCite* by Cohan and colleagues (2019) marks a significant advancement in the field. This dataset spans both computer science and medicine related domains, featuring about 11,000 manually annotated citation contexts divided into three broader categories. This cross-domain coverage makes SciCite more versatile and positions it as a valuable benchmark for CIC.

The main problem of these two datasets is represented by their imbalanced distribution (**Table 1**), which is a major issue when trying to solve the CIC task with automated methods (Kunnath et al., 2021). Previous works tried to overcome the skewness of datasets for CIC by applying SMOTE (Jurgens et al., 2018; Nazir et al., 2020; Qayyum & Afzal, 2019) or by re-balancing the original corpus distribution (Dong & Schafer, 2011). Our approach, in contrast, targets each class individually, thereby facilitating a focused and independent treatment of each intent category. This design enables our classifiers to effectively address even underrepresented classes, with the explicit aim of mitigating the challenges posed by data imbalance.

2.3 Historical Approaches and Promising PLMs for CIC

The evolution of methodologies for addressing the CIC task mirrors advancements in machine learning (ML) and natural language processing (NLP). Early rule-based systems (Garzone & Mercer, 2000; Nanba et al., 2011) relied on handcrafted heuristics but exhibited limited generalizability due to their dependence on predefined patterns (Kunnath et al., 2021), resulting in rigid frameworks. Traditional ML approaches, such as feature engineering and predictive modeling (Agarwal et al., 2010; Bakhti et al., 2018; Dong & Schafer, 2011; Hassan et al., 2017; Jurgens et al., 2018; Pranckevičius & Marcinkevičius, 2017; Pride & Knoth, 2017; Sula & Miller, 2014; Teufel et al., 2006; Valenzuela-Escárcega et al., 2015; Xu et al., 2013), reduced reliance on manual rules but still required human expertise for feature selection (Su et al., 2019), yielding suboptimal results.

The advent of deep learning (DL) introduced neural architectures like CNNs and RNNs, automating feature extraction and improving sequence processing. Advanced variants, including LSTMs (Hochreiter & Schmidhuber, 1997; Graves, 2013) and GRUs (Cho et al., 2014), addressed challenges such as vanishing gradients (Bengio et al., 1994; Pascanu et al., 2013) and short-term memory constraints (Shiri et al., 2023), enabling better capture of long-term dependencies (Dutta et al., 2020). Bidirectional LSTMs (BiLSTMs) further enhanced contextual analysis by processing sequences in both directions (Aldhyani & Alkahtani, 2021; Cornegruta et al., 2016). While empirical comparisons between GRUs and LSTMs yield mixed results (Chung et al., 2014; Dutta et al., 2020), both architectures outperform traditional methods in handling long text sequences, advancing CIC performances.

Transformers (Vaswani et al., 2017) revolutionized NLP by replacing recurrence with self-attention, enabling parallel processing and superior context modeling, thereby resulting in the proficient processing of sequential text data and in models capable of performing complex textual tasks (W. X. Zhao et al., 2023). Pretrained language models (PLMs), like *BERT* (Devlin et al., 2019), leveraged transformer architectures and large-scale training to achieve *state-of-the-art* (SOTA) results across NLP tasks.

For CIC, domain-specific PLMs became critical due to the specialized lexicon and syntax of academic texts. *SciBERT* (Beltagy et al., 2019), specifically pretrained on scientific corpora, outperforms general-purpose models by aligning with academic language structures through its specialized *SciVOCAB* vocabulary (Cohan et al., 2019). Its applications in CIC, including prompt-based frameworks such as CitePrompt (Lahiri et al., 2023), demonstrated robust performances, in particular on the ACL-ARC dataset on which the authors obtained SOTA results. Moving on from BERT and its derivations, which employed a masked language modelling objective, other training strategies and objectives resulted beneficial in understanding citation contexts despite the use of general training corpora. *XLNet* (Yang et al., 2019) is an example of such models, it combines autoregressive and autoencoding objectives, further improving the ability to understand citation contexts without specialized vocabularies. A fine-tuned version of XLNet, *ImpactCite* (Mercier et al., 2021), achieved SOTA results on SciCite, while it was never applied to solve the CIC task on ACL-ARC.

Thus, while SciBERT's domain-specific pretraining makes it particularly effective for processing academic texts, XLNet's generalized pretraining and advanced language modeling objectives exemplify the trade-offs between specialization and adaptability, as evidenced by the findings presented in this section. The studies discussed here, which leverage these two PLMs, primarily focus on classifying raw citation contexts without expanding their semantic scope. Furthermore, these approaches tend to prioritize performance metrics, often neglecting the underlying nature of the classifications produced by the models. This emphasis on performance can result in reduced interpretability regarding the mechanisms and rationale behind specific predictions, underscoring the need for methodologies that balance accuracy with a deeper understanding of model behavior.

Additionally, given the distinct strengths of XLNet and SciBERT, we argue and demonstrate in this work how their integration can yield robust and reliable performance by capturing complementary aspects of citation contexts. By combining these architectures, we aim to highlight their respective capabilities and leverage their synergies to address diverse challenges in citation intent classification effectively.

2.4 Ensemble Strategies and Explainable AI for Interpretable Classification

An interesting research direction in classification tasks is related to ensemble strategies (ES), in which a set of weak - or baseline - learners are employed by means of an aggregation function to produce a single output through the different predictions of the base models (Mohammed & Kora, 2023). ES are defined by two components: (1) *baseline models*, which can be homogeneous (same type) or heterogeneous (different types), and (2) *aggregation methods*, ranging from simple voting strategies, such as *max voting* (Kim et al., 2003), *average voting* (Montgomery et al., 2012), or *weighted average voting* (Latif-Shabgahi, 2004), to meta-learning approaches where a secondary model learns from base predictions (Mohammed & Kora, 2023; Soares et al., 2004). In particular, the three most common ES are known as *Bagging* (Breiman, 1996), *Boosting* (Freund & Schapire, 1996), and *Stacking* (Smyth & Wolpert, 1997). In this study, we focus on Stacking and some of its variants, driven by the objective of harnessing the complementary strengths of the base PLMs utilized in our framework.

In *Stacking*, multiple base models (sometimes referred to as *level-0* models) of the same or of different types are trained on the same or on different subsets of data. The predictions of these base models, either in the form of probabilities or class labels, are then combined and used to train a meta-model (sometimes referred to as *level-1* model), or aggregated through a more traditional voting strategy.

StackingC (Seewald, 2002) is a notable variant of Stacking, in which *Multiple Linear Regression* (MLR)¹ is used to predict per-class specific probabilities from each set of level-0 models, reducing the complexity of the overall aggregation function, and leading to better performances in multiclass settings. Another alternative to traditional stacking is the *Geometric Framework* described by Wu and colleagues (2023), in which MLR is used to minimize the Euclidean Distance (ED) from the predicted and the ideal points in n -dimensional spaces, thereby finding optimal weights at dataset level to apply to base models predictions as weighting schemes for weighted voting strategies.

ES demonstrated to improve classification performances in a wide range of tasks from various domains (Mohammed & Kora, 2023), and in many cases they perform better than more traditional methods when dealing with imbalanced classification tasks. Indeed, by harnessing the power of multiple classifiers, it is possible to deal more efficiently with underrepresented classes, as demonstrated in recent studies (Khan et al., 2024; L. Liu et al., 2022; S. Liu et al., 2017; D. Zhao et al., 2021). Within ES it is also possible to harness the power of multiple baseline PLMs to produce different outputs to fuse through an aggregator (Huang et al., 2024; Jiang et al., 2023). To build such a framework there are various possibilities which may involve differentiated folds of data, heterogeneous - or homogeneous - baseline PLMs, and multiple learning stages (Monteiro et al., 2021).

However, the complexity of ES and the opacity of PLMs working dynamics (e.g., latent representations in transformer layers) challenge interpretability (Longo et al., 2024). *Explainable AI* (XAI) tries to address this by making classification processes more transparent, and this is critical for trust in tasks like CIC. XAI techniques such as *SHAP* (*SHapley Additive explanations*) (Lundberg & Lee, 2017), which quantifies features contributions using Shapley values, provide some useful insights into predictions, models, and ultimately the datasets employed.

Furthermore, identifying model and class-specific relevant features as token-level contributions could also help in providing different viewpoints on the inherent semantic of various citation intents, adding building blocks to the discourse on CIC. Thereby, XAI methods not only help in demystifying "*black-box*" outputs (Ribeiro et al., 2016) but also guide model and dataset refinements, enhancing reliability for downstream applications, as it has been observed in other domains (Leichtmann et al., 2023; Trindade Neves et al., 2024). By coupling ensemble robustness with XAI transparency, it is possible to achieve a balance between accuracy and interpretability – essential for advancing reproducible and trustworthy AI systems in academic and industrial settings.

¹ Also denoted as *Multi-Response Linear Regression* (Seewald, 2002).

3 Models and Experiments

This section outlines the implementation and training dynamics of *CiteFusion*, the Ensemble Strategy (ES) designed to enhance the performances, as measured by Macro-F1² and accuracy scores, in the classification of citation intents on both SciCite and ACL-ARC datasets.

3.1 Datasets, Mapping, and PLMs

For our experiments, we utilized the two datasets introduced in the previous section: *SciCite* and *ACL-ARC*³. As noted, these datasets exhibit significant class imbalance, a characteristic that is consistently reflected across all three predefined splits (*training*, *validation*, and *test*). This ensures that the imbalances inherent in the datasets are preserved and appropriately accounted for throughout the experimental process. To enhance the interoperability of the classifications performed by our models, we designed two mapping schemes that translate the original labels of both datasets into object properties selected from CiTO (Peroni & Shotton, 2012). These mappings are summarized in *Table 2*.

Dataset	Original Labels	Mapping
SciCite	Method	http://purl.org/spar/cito/usesMethodIn
	Background	http://purl.org/spar/cito/obtainsBackgroundFrom
	Result	http://purl.org/spar/cito/usesConclusionsFrom
ACL-ARC	Background	http://purl.org/spar/cito/obtainsBackgroundFrom
	Uses	http://purl.org/spar/cito/usesMethodIn
	CompareOrContrast	http://purl.org/spar/cito/discusses
	Extends	http://purl.org/spar/cito/extends
	Motivation	http://purl.org/spar/cito/obtainsSupportFrom
	Future	http://purl.org/spar/cito/citesAsPotentialSolution

Table 2. Mapping of the original schemes of *SciCite* and *ACL-ARC* datasets, with object properties from CiTO.

In developing *CiteFusion*, we decided to leverage the Pretrained Language Model (PLM) architectures outlined in *Section 2.3*: *SciBERT* (Cased) (Beltagy et al., 2019) and *XLNet* (base-Cased) (Yang et al., 2019). The decision to employ both architectures as individual components was driven by the aim to harness the complementary strengths of domain-specific

specialization and generalized adaptability. By integrating both models, we sought to achieve a balance between precision in academic language processing and robust generalization across diverse citation contexts.

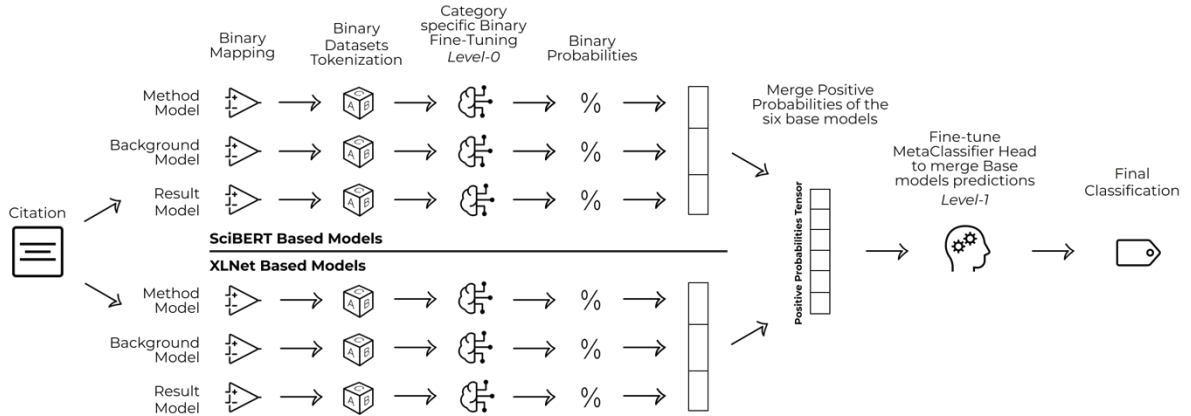


Figure 2. Overview of the ES described in this work. The figure shows the working dynamics, and the steps followed to produce the final Ensemble Classifiers for the *SciCite* dataset. The strategy follows the same exact procedure in classifying Citations from *ACL-ARC* but, instead of 6 base PLMs, it employs 12 of them (two for each category).

² Macro-F1 is the reference metric used to compare the results obtained in CIC.

³ While we utilized *SciCite* from the official Huggingface release (<https://huggingface.co/datasets/allenai/scicite>), we took the preprocessed *ACL-ARC* dataset from <https://github.com/allenai/scicite> (Cohan et al., 2019).

3.2 Ensemble Strategy

As outlined in the *Introduction*, to address the CIC task, we designed *CiteFusion* (exemplified in **Figure 2**), an ensemble strategy that leverages two different PLM architectures and fuse their respective strengths. This approach was applied to both SciCite and ACL-ARC datasets. The adoption of CiteFusion was primarily motivated by two key factors: (1) the relatively limited number of data points in both datasets which, together with (2) the significant class imbalance present in the data, restricts the ability to fully harness the potential of a single PLM for this task. These challenges suggested the use of a stacked ensemble approach to effectively address and mitigate their impact.

Problem Definition

To provide a clear and concise description of the ES for the two datasets, we briefly formalize its general structure. Let's denote with K_D the number of classes $CL_D = \{c_{1D}, c_{2D}, \dots, c_{K_D}\}$ of a dataset D , and with M_D the number of models used as level-0 predictors for the same dataset. Each dataset is represented as $D = \{(s_i, x_i, y_i)\}_{i=1}^N$, where:

- x_i : a sentence containing a citation context;
- s_i : the title of the section in which x_i is contained;
- $y_i \in CL_D$: the corresponding citation intent label.

Now we can define the input sentence β to be classified in two different settings:

- $\beta_{WS} = s_i + ". " + x_i$: input sentence **with** section title (WS setting)⁴;
- $\beta_{WoS} = x_i$: input sentence **without** section title (WoS setting).

The use of two different input sentence structures is functional to highlight the role of section titles to investigate how the classification task changes when extended context is provided⁵. For each dataset, we separately investigate the role of section titles by training an independent ensemble classifier for WS and WoS settings.

Base Classification (Level-0)

To address the two challenges previously described and thereby maximize the efficacy of our models in a data-scarce and imbalanced setting, we transform the original multi-class problem into multiple binary tasks. Thereby, for each class $c_k \in CL_D$, we define a binary dataset D_k as $D_k = \{(s_i, x_i, z_i^k)\}_{i=1}^N$, where:

$$z_i^k = \begin{cases} 1, & \text{if } y_i = c_k \\ 0, & \text{otherwise.} \end{cases}$$

For SciCite this transformation leads to $\#D_{K_{SciCite}} = 3$ binary datasets ($k_{SciCite} = 3 = [0, 1, 2]$), while for ACL-ARC we obtain $\#D_{K_{ACLARC}} = 6$ binary datasets ($k_{ACLARC} = 6 = [0, 1, 2, 3, 4, 5]$).

As stated before, we employ two distinct *PLM* architectures: $SB = SciBERT$ and $XN = XLNet$. Then, for each binary task we fine-tune both PLMs independently, resulting in $M_D = 2 \cdot K_D$ classifiers.

For the two datasets we thereby obtain:

- $M_{SciCite} = 2 \cdot K_{SciCite} = 6$ binary classifiers for SciCite;
- $M_{ACLARC} = 2 \cdot K_{ACLARC} = 12$ binary classifiers for ACL-ARC.

⁴ The + sign represents a concatenation of strings in this context.

⁵ Examples of the same input sentence from SciCite in both WS and WoS settings (in bold we highlighted s_i):

- Example of β_{WS} : "**Introduction**. However, how frataxin interacts with the Fe-S cluster biosynthesis components remains unclear as direct one-to-one interactions with each component were reported (IscS [12,22], IscU/Isc1 [6,11,16] or ISD11/Isc11 [14,15])."
- Example of β_{WoS} : "However, how frataxin interacts with the Fe-S cluster biosynthesis components remains unclear as direct one-to-one interactions with each component were reported (IscS [12,22], IscU/Isc1 [6,11,16] or ISD11/Isc11 [14,15])."

Let $M_{\theta_k}^{PLM}$ denote the k -th binary classifier parametrized by θ_k after fine-tuning on the respective D_k , where $PLM \in \{SB, XN\}$. Each model $M_{\theta_k}^{PLM}$ outputs a positive probability \hat{p}_k^{PLM} denoting the degree of confidence of the classifier for that input sentence β to belong to class k . In this way, each model outputs a probability score for each input sentence of the original dataset D in a binary context.

The outputs of all binary classifiers for each input sentence β_i , with $i \in N$, are then concatenated to form N_D instances of P_D , being each P_D a M_D -dimensional representation of the ensemble prediction at level-0. Specifically, we obtain:

- $P_{SciCite_i} = [\hat{p}_0^{SB}, \hat{p}_1^{SB}, \hat{p}_2^{SB}, \hat{p}_0^{XN}, \hat{p}_1^{XN}, \hat{p}_2^{XN}]$ for each $i \in N_{SciCite}$;
- $P_{ACLARC_i} = [\hat{p}_0^{SB}, \hat{p}_1^{SB}, \hat{p}_2^{SB}, \hat{p}_3^{SB}, \hat{p}_4^{SB}, \hat{p}_5^{SB}, \hat{p}_0^{XN}, \hat{p}_1^{XN}, \hat{p}_2^{XN}, \hat{p}_3^{XN}, \hat{p}_4^{XN}, \hat{p}_5^{XN}]$ for each $i \in N_{ACLARC}$.

Meta-Classification (Level-1)

The level-0 vectors serve as input to various meta-classification strategies, which aim to map the concatenated binary predictions back to the original multi-class space. In this context, we explored both supervised and unsupervised approaches, starting with traditional voting mechanisms and progressing to more advanced techniques.

We first evaluated three baseline voting strategies, adapted to the binary setting previously described. Knowing that the K_D predictions from SciBERT PLMs for a single instance $i \in N$ of dataset D are $P_{D_i}^{SB} = [\hat{p}_{1_i}^{SB}, \hat{p}_{2_i}^{SB}, \dots, \hat{p}_{k_i}^{SB}]$ and that the corresponding predictions for the same instance from XLNet are $P_{D_i}^{XN} = [\hat{p}_{1_i}^{XN}, \hat{p}_{2_i}^{XN}, \dots, \hat{p}_{k_i}^{XN}]$, we defined:

1. **Max Voting:** The final prediction for each P_{D_i} is determined by selecting the class with the highest prediction score across all binary classifiers. Thereby, for each P_{D_i} we obtain the final prediction⁶ \hat{y}_i from the ensemble classifier (EC) as $\hat{y}_i = \max(P_{D_i})$.
2. **Average Voting:** In each instance of P_{D_i} , the predictions from the two PLM architectures are averaged for each class, and the class with the highest average score is selected. we obtain:

$$\hat{y}_i = \max\left(\frac{P_{D_i}^{SB} + P_{D_i}^{XN}}{2}\right), \forall i \in N_D$$

3. **Majority Voting:** Each model votes for a class only if its predicted probability exceeds a predefined threshold. The final prediction is determined by the class with the most votes, with tie-breaking rules applied when necessary. Formally, we apply a threshold $T = 0.5$ (this value was selected because each level-0 prediction $\hat{p}_k^{PLM} \in P_D$ is in range $[0, 1]$) to determine binary votes, in a way in which the vote is considered 1 when $\hat{p}_k^{PLM} \geq T$, and 0 otherwise. Then we combine SciBERT and XLNet votes for each class and identify the class with the maximum number of votes. If we have more than one class with the highest number of votes for a particular instance, we apply an average voting to choose between them.

⁶ The predicted class label \hat{y} from all voting strategies corresponds to the index of the highest prediction probability. For simplicity, we omit explicitly writing the *argmax* operation in the following descriptions. However, it should be understood as follows:

- When dealing with K_D predictions (e.g., in average voting), the *argmax* operation is applied directly over the combined probabilities to determine the final class label.
- When dealing with M_D predictions (e.g., in max voting), two separate *argmax* operations are performed. Specifically, we first compute $MAX_{SB} = \max(P_{D_i}^{SB})$ and $MAX_{XN} = \max(P_{D_i}^{XN})$. These values are then compared and the index of the higher score determines the predicted class.

Building on these traditional strategies, we also adapted the geometric framework described by Wu and colleagues (2023) to our case, as introduced before (see *Section 2.4*). We mapped each binary label to a 2-dimensional space, where the predictions from the two PLM architectures for the same class are represented as a pair of coordinates. Formally, for each class c_k , the prediction couple is defined as $v_k = [\hat{p}_k^{SB}, \hat{p}_k^{XN}]$. The geometric framework operates by minimizing the Euclidean distance (ED) between the prediction vector v_k and the target representation⁷ in the 2-dimensional space. Specifically, the objective is to find the optimal weights that aligns the predicted vectors with their corresponding true labels, thereby obtaining a weight for each model of each couple. This is achieved with Multiple Linear Regression (MLR), through which we extract two weights for each class – one for each architecture –, obtained as a result of the minimization of the ED between the ideal and the predicted points in the validation split of D . For each class c_k we thereby obtain two weights $W_k = [w_k^{SB}, w_k^{XN}]$, and we normalize them to sum to 1. Therefore, for the resulting weights of each couple we have $w_k^{SB} + w_k^{XN} = 1$. This weighting scheme is therefore computed at dataset level for each class, and it is then applied to each level-0 prediction vector of the test set of D . Finally, we applied the three voting schemes described before to these weighted predictions to retrieve the resulting multiclass labels.

The final voting strategy we employed is *StackingC* (Seewald, 2002), which utilizes MLR to predict class-specific probabilities from pairs of level-0 predictions. To adapt this method to our ensemble, we extracted the prediction vectors $v_k = [\hat{p}_k^{SB}, \hat{p}_k^{XN}]$ for each class c_k from the validation split of the dataset. These vectors, representing the outputs of the two PLM architectures for each class, were used to train a MLR model. The MLR model learns to map these paired predictions to binary targets, effectively determining the probability of each class in the context of the specific classifier pair. Once trained, the MLR model generates probability estimates for all classes, which are then converted into probabilities, and the final prediction is therefore obtained through an *argmax* over these probabilities⁸.

Beyond voting-based approaches, we implemented machine learning related algorithms to use as meta-classifiers. In this context, we performed a Grid Search over possible hyperparameter combinations for Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Finally, we implemented a Feed Forward Neural Network, which takes the concatenated level-0 predictions as input and learns to map them to the final multi-class predictions. This approach allows the system to capture complex interactions between the binary predictions, ideally improving classification accuracy.

3.2.1 Training Dynamics and Computational Instability

To employ the baseline models for the binary task we developed a fine-tuning loop structured to overcome the overfitting problem that usually arises when adapting LMs to downstream tasks. This loop allows for detailed assessments of the model's performance on the validation set, and it is accompanied by a scheduler to decrease learning rate on validation loss plateaus. Specifically, the model is evaluated every 10 batches of training data within each epoch. During each evaluation, the model's performance in terms of validation loss is compared to the best performance recorded up to that point and the best model's state in terms of validation loss is saved as a checkpoint. This process ensures that the model is continuously monitored, and the best performing version is preserved and finally retrieved at the end of the loop. Additionally, an early stopping mechanism has been implemented to stop the fine-tuning process after 50 evaluations without performance increase.

⁷ i.e. the target label y_k .

⁸ Even though the settings and tools of the geometric framework and of *StackingC* seems similar, we would like to highlight that these two methods, also when adapted to our context, are different. In particular, the geometric framework serves to align the class-specific predicted probabilities (v_k) with their true labels in a geometric space by learning class-specific weights for each model. In this context, MLR is used to minimize the ED between the two (predicted and ideal) 2-dimensional points through the learned coefficients. The learned weights are then applied to test predictions, and the final label is computed through the original voting strategies. In contrast, *StackingC* uses MLR to directly predict class probabilities by treating the two PLM predictions as input features to a meta-model. In this case, we use the MLR's output probabilities directly, selecting the highest one. In conclusion, while *StackingC* represents a stacking ensemble, the geometric framework serves as a custom weighting scheme.

Level-0 models were fine-tuned using varying learning rates and weight decay values across experiments⁹. We used the *AdamW* optimizer with *cross-entropy loss*, applying weight decay to all parameters except for *biases* and *LayerNormalization* terms. Excluding *LayerNormalization* weights from decay is crucial because they are part of a normalization mechanism, and applying decay could disrupt the normalization process, destabilizing training dynamics. Similarly, bias terms were excluded from weight decay since penalizing them is unnecessary and can hinder model performance by interfering with data fitting. Empirical evidence shows that excluding these parameters from the weight decay application improves both training stability and effectiveness (Ba et al., 2016; Lahiri et al., 2023; Loshchilov & Hutter, 2017). Mixed-precision training was also employed, utilizing both 16-bit and 32-bit floating-point numbers to reduce memory usage and accelerate computations without compromising the performance of the models. To maintain coherence with the ensemble strategy, we do not evaluated level-0 models as individual classifiers on the test sets of binary datasets employed for their fine-tuning.

All experiments were conducted using a fixed seed and deterministic algorithms to ensure reproducibility. However, complete reproducibility is not always guaranteed, particularly in PyTorch when using mixed-precision training. The use of 16-bit floating-point numbers can introduce small numerical differences in computations, which may accumulate over time, leading to slight variations in model parameters across training runs. While these differences are generally minor, they can impact reproducibility in fine-grained evaluations and when working with highly sensitive models, such as those used as level-0 classifiers in our stacked ensembles. Additionally, as reported in the PyTorch documentation, *fully reproducible results are not assured across different PyTorch releases, individual commits, or platforms [...] and results may differ between CPU and GPU executions, even with identical seeds*. Furthermore, in dealing with imbalanced datasets, reproducibility is not always guaranteed.

To address these reproducibility challenges, we conducted computational instability analyses on the best-performing ensemble classifiers derived from this study. Specifically, for both datasets, we repeated the fine-tuning process of the level-0 models and applied (trained) the same meta-classification strategy¹⁰ across 10 additional and independent experimental runs. Throughout these repetitions, we kept all hyperparameters and settings consistent to ensure a fair and controlled evaluation. The primary objectives of these analyses were twofold: (1) to assess the robustness of our ES when applied to imbalanced datasets, and (2) to investigate whether mixed-precision training and fine-grained evaluations significantly impact performance across different runs and thereby reproducibility. The results of these analyses are reported in *Section 4*.

3.2.2 Explainers

Following the development of CiteFusion to address the CIC task, we conducted a series of experiments using SHAP (Lundberg & Lee, 2017) to enhance the interpretability and trustworthiness of our classifiers. To ensure consistency, SHAP was applied systematically across all the ECs, and SHAP values were computed at both levels (level-0 and level-1) of the ensemble framework for both datasets, SciCite and ACL-ARC. At level-0, SHAP was employed to explain the contributions of individual tokens, considered as features influencing the predictions made by the base models. The analysis of the results of these experiments provides insights into how specific words or phrases impact classification outcomes in binary settings, also highlighting the differences between the two PLM architectures employed.

At level-1, SHAP was instead utilized to elucidate the aggregation process of base model predictions within the FFNN metaclassifier. This approach facilitates an understanding of which base models contribute most significantly to shaping the final predictions for each class. By applying SHAP at both levels, we gain a comprehensive view of the classification dynamics of our ES while also uncovering dataset-specific characteristics inherent to both SciCite and ACL-ARC. In this context, SHAP serves as a valuable tool for enhancing transparency and fostering a deeper understanding of models' behavior and underlying data properties.

⁹ To inspect the full hyperparameters setting refer to the publicly available code (Paolini, 2024d).

¹⁰ The best performing for each dataset in WS settings.

4 Experimental Results and Discussion

This section addresses critical inquiries related to the performance, robustness, and interpretability of CiteFusion when applied to both SciCite and ACL-ARC datasets. The impact of incorporating section titles into input sentences on classification performance is examined, also focusing on how the integration of this structural element (utilized as semantic feature) reshapes the way in which base models perceive citation contexts, highlighting key differences between the different kind of features considered by the two Pretrained Language Model (PLM) architectures when performing binary classifications. The effect of these perceptual shifts is also examined for the FFNNs metaclassifiers, detailing the contributions of each model for each class-specific output of both datasets.

The robustness of our binarized couple-based ensemble strategy is also assessed to evaluate its consistency under imbalanced data settings. The same analyses reveal the effects of mixed precision and fine-grained evaluations on reproducibility, to assess both the reliability and replicability of our findings. Additionally, the classification mechanisms of our Ensemble Classifiers (ECs), analyzed through features contributions at both level-0 and level-1, enhance the interpretability of CiteFusion by clarifying some dynamics of the stacked processes. Finally, analyses on some classification errors made by our models are presented. These efforts aim to provide a comprehensive understanding of the factors driving the effectiveness of our approach.

4.1 Results

Experimental results revealed a positive influence of incorporating section titles. Across all experiments conducted on both datasets, the WS (with section titles) setting consistently outperformed the WoS (without section titles) setting. For SciCite, the mean improvements across all the aggregation functions and models utilized in the experiments were $\Delta_A = 0.73$ and $\Delta_{MF1} = 0.93$. Similarly, the mean improvements w.r.t. the final classification scores for ACL-ARC were $\Delta_A = 1.86$ and $\Delta_{MF1} = 1.55$. Furthermore, as evident from **Table 3**, FFNNs consistently outperformed all other aggregators, except for the WoS setting in ACL-ARC, in which Random Forest (RF) performed better.

Setting	Aggregation Function	SciCite		ACL-ARC	
		Accuracy	Macro-F1	Accuracy	Macro-F1
WS	Max	89.94	88.51	79.86	71.90
	Avg	90.16	88.98	78.42	68.58
	Maj	90.10	88.92	78.42	68.58
	W-Max	89.99	88.56	79.14	71.33
	W-Avg	90.21	88.99	78.42	68.35
	W-Maj	90.05	88.74	78.42	70.28
	StackingC	90.10	88.81	82.73	73.91
	RF	88.76	87.36	80.58	73.02
	SVM	89.89	88.75	78.42	68.16
	LR	90.16	89.08	79.14	70.86
	KNN	88.70	87.25	79.86	69.20
	FFNN	90.08	89.60	81.29	76.24
	Mean	89.85	88.63	79.62	70.81
WoS	Max	88.81	87.30	76.98	69.68
	Avg	89.46	88.04	77.70	69.42
	Maj	89.46	88.04	78.42	70.10
	W-Max	88.92	87.67	77.70	69.78
	W-Avg	89.46	88.11	78.42	67.74
	W-Maj	89.46	88.11	79.14	68.38
	StackingC	89.24	87.91	76.98	62.15
	RF	88.54	86.97	79.86	72.44
	SVM	89.24	87.75	74.82	69.76
	LR	89.13	87.72	79.14	71.03
	KNN	88.17	86.56	76.26	68.42
	FFNN	89.56	88.22	76.98	71.46
	Mean	89.12	87.70	77.76	69.26

Table 3. The table presents Accuracy and Macro-F1 scores for different aggregation strategies of level-0 models on both SciCite and ACL-ARC datasets. It includes mean results across all level-1 models (heads) for both WS and WoS settings. Individual scores are reported for three voting strategies – Max, Average (Avg), and Majority (Maj) – along with their weighted versions (W-Max, W-Avg, W-Maj). Scores for machine learning (ML) algorithms are also provided. For FFNN, the reported scores reflect the best results from 10 runs of CiteFusion in the WS setting, and single-run results in the WoS setting.

The best-performing model for the SciCite dataset was trained in WS setting. Specifically, it attains a Macro-F1 score of 89.60%, a Micro-F1 and accuracy scores of 90.80%, and a weighted F1 score of 90.87%. Additionally, it demonstrates a precision (macro average) of 88.57% and a recall (macro average) of 90.95%. The class-specific performance metrics are as follows: for the *Method* class, accuracy and F1 scores are 89.60% and 91.65%, respectively; for the *Background* class, 91.06% and 91.85%; and for the *Result* class, 91.89% and 85.30%.

For the ACL-ARC dataset, the best-performing model is again trained in WS setting. This achieves a Macro-F1 score of 76.24%, a Micro-F1 and accuracy scores of 81.29%, and a weighted F1 score of 80.77%. The precision (macro average) is 80.91%, and the recall (macro average) is 72.84%. Class-specific results are as follows: for the *Background* class, accuracy and F1 scores are 92.96% and 86.84%, respectively; for the *Uses* class, 69.23% and 73.47%; for the *CompareOrContrast* class, 72.00% and 78.26%; for the *Extends* class, 80.00% and 88.89%; for the *Motivation* class, 42.86% and 50.00%; and for the *Future* class, 80.00% and 80.00%.

As previously mentioned, we conducted additional analyses to assess the computational instability of our Ensemble Strategy (ES). These analyses were carried out exclusively in the WS setting¹¹ repeating the same experiment 10 additional times for both ACL-ARC and SciCite. For both the analyses, we employed the same FFNN architecture as the head for the level-0 models. The results of these investigations are summarized in **Table 4**.

Setting	Run	SciCite		ACL-ARC	
		Accuracy	Macro-F1	Accuracy	Macro-F1
WS	0	90.32	89.01	80.58	75.28
	1	90.53	89.39	80.58	73.19
	2	90.64	89.50	79.14	72.12
	3	90.53	89.35	79.86	73.35
	4	90.59	89.42	79.86	70.76
	5	90.32	89.01	79.86	72.81
	6	90.80	89.60	79.86	72.36
	7	90.64	89.50	81.29	76.24
	8	90.59	89.38	80.58	73.91
	9	90.59	89.43	81.29	75.22
	10	90.37	89.22	82.01	75.54
<i>Mean (Std)</i>		90.53 (0.14)	89.34 (0.18)	80.44 (0.79)	73.71 (1.61)

Table 4. Results of computational instability analyses on both datasets. Accuracy and Macro-F1 scores of each run are reported, together with score and dataset specific means and standard deviations (grey row). Run 10 represents the first (base) experiment, while runs from 0 to 9 represent repeated runs of it.

Finally, for plots detailing the results involving SHAP to explain the predictions of our ECs, we refer the reader to **Figures A.1, A.2, A.3, A.4, A.5, and A.6** in **Appendix A.1** and to the following sections. The key insights derived from this technique reveal the most influential tokens for level-0 binary predictions and identify the most impactful models for level-1 metaclassification via FFNNs. The paper will present the most interesting findings related to explainability while discussing them in **Section 4.3**.

4.2 Discussion

CiteFusion (WS) models surpass the previous state-of-the-art (SOTA) results in both SciCite and ACL-ARC benchmarks, as depicted in **Tables 5 and 6**. Furthermore, the FFNN metaclassifier head consistently outperforms the previous SOTA models for both SciCite and ACL-ARC across all the computational instability runs.

Additionally, the inclusion of section titles in input sentences consistently improves the performance of ECs in CIC, as evidenced by gains in both accuracy and macro-F1 scores across the datasets w.r.t. WoS setting. Nevertheless, even in WoS setting, the robust performance of the ECs highlights the effectiveness of CiteFusion in dealing with underrepresented classes in both datasets. By training complementary couples of level-0 models on binary tasks, we enabled a focused approach to handling each class, effectively mitigating class imbalance.

¹¹ Due to limited computational resources.

The results are summarized in **Table 5** for SciCite, and in **Table 6** for the ACL-ARC dataset. These two tables compare the scores of our models (*CiteFusion*) against the two benchmarks¹². In **Table 5** the comparison is against the two SOTA models for SciCite: *ImpactCite* (Mercier et al., 2021) and *CitePrompt* (Lahiri et al., 2023).

Model	Per-Class Accuracy			Per-Class F1 Scores			Accuracy	Precision	Recall	Micro-F1	Weighted F1	Macro-F1
	MET	BKG	RES	MET	BKG	RES						
<i>CiteFusion (WS)</i>	94.67	91.34	95.59	91.65	91.85	85.30	90.80	88.57	90.95	90.80	90.87	89.60
<i>CiteFusion (WoS)</i>	93.65	89.89	95.59	89.58	91.00	84.86	89.73	89.77	87.40	89.73	89.68	88.48
<i>ImpactCite</i> ¹³	85.79	88.34	92.67	87.00	90.00	85.00	88	NA	NA	88.13	88	88.93
<i>CitePrompt</i>	NA	NA	NA	NA	NA	NA	87.56	NA	NA	NA	NA	86.33

Table 5. Results for the SciCite dataset: This table compares our models, *CiteFusion* (in both WS and WoS settings), with the current state-of-the-art models for the CIC task on this dataset.

In **Table 6** the comparison is against *CitePrompt* (the current best-performing model for this task), and the model obtained by Cohan and colleagues (2019), which is presented as *Structural Scaffolds* in the table.

Model	Background		Uses		ComOrCon		Extends		Motivation		Future		A	P	R	MiF1	W-F1	MaF1
	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1						
<i>CiteFusion (WS)</i>	85.61	86.84	90.65	73.47	92.81	78.26	99.28	88.89	95.68	50.00	98.56	80.00	81.29	80.91	72.84	81.29	80.77	76.24
<i>CiteFusion (WoS)</i>	81.29	83.33	89.93	72.00	89.21	63.41	97.84	66.67	97.12	60.00	98.56	83.33	76.98	79.86	69.27	76.98	75.86	71.46
<i>Structural Scaff.</i>	NA	83.5	NA	75.0	NA	71.1	NA	66.7	NA	44.4	NA	66.7	NA	81.3	62.5	NA	NA	67.9
<i>CitePrompt</i>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	78.42	NA	NA	NA	NA	68.39

Table 6. Results for the ACL-ARC Dataset: This table compares our proposed models, *CiteFusion* (in both WS and WoS settings), with the current state-of-the-art models for the CIC task on this dataset (note that *Structural Scaffolds* name is abbreviated). For clarity, we use abbreviations to represent evaluation metrics. Evaluation scores are denoted as follows: *A* for Accuracy, *P* for Precision, *R* for Recall, *MiF1* for Micro-F1, *W-F1* for Weighted-F1, and *MaF1* for Macro-F1.

These tables demonstrate the superior performance of our strategy compared to current state-of-the-art models, underscoring its effectiveness in handling imbalanced datasets. This is especially evident when examining the significant improvements achieved by *CiteFusion* (WS) on the ACL-ARC dataset, with nearly an 8% increase in Macro-F1 and a 3% rise in accuracy compared to the previous SOTA. These results, together with the scores obtained on the SciCite dataset, highlight our strategy's ability to accurately classify and discern even highly underrepresented classes.

Model	Method		Background		Result	
	P	R	P	R	P	R
<i>CiteFusion (WS)</i>	93.46	89.90	92.65	91.06	79.60	91.86
<i>CiteFusion (WoS)</i>	92.04	88.08	89.22	92.27	85.26	82.63

Table 7. Class-specific Precision (P) and Recall (R) scores for the SciCite Dataset.

Furthermore, we report in **Tables 7 and 8** class-specific precision and recall scores for SciCite and ACL-ARC, respectively, in both WS and WoS settings. The integration of section titles in SciCite yields mixed yet insightful performance differences across classes. For the underrepresented *Result* class, *CiteFusion* in WS setting achieves a recall of 91.86% (+9.23% over WoS), but at the expense of lower precision, which drops to 79.60% (from 85.26% in WoS). This indicates that section titles assist in capturing more instances of the *Result* class but introduce false positives. This behavior aligns with expectations, as section headers such as "Results" likely provide contextual cues that enhance the detection of infrequent *Result* citations, prioritizing recall, but eventually introducing biases. For the majority classes, *Method* and *Background*, *CiteFusion* in WS setting demonstrates improved precision, achieving 93.46% compared to 92.04% for *Method*, and 92.65% compared to 89.22% for *Background*, while increasing the recall from 88.08% to 89.90% for the *Method* class, and decreasing it from 92.27% to 91.06% in the *Background* class. These results suggest that section titles help refine the model's focus without compromising coverage for these classes.

¹² The tables, for *CiteFusion*, report the best scores obtained through 10 runs of the ECs in WS settings, and the scores obtained through a single run in WoS settings. The other models reported in these tables are the current state-of-the-art models for the two datasets in CIC.

¹³ Even if reference scores of *ImpactCite* (Mercier et al., 2021) for Accuracy, Weighted-F1, and class-specific F1s are not directly reported in the paper, they are present in the code released by the authors, publicly available at https://github.com/DominiqueMercier/ImpactCite/blob/main/ImpactCite_Intent/Tester.ipynb.

In ACL-ARC, the inclusion of section titles has a significant but uneven impact on underrepresented classes. For the *Extends* class, CiteFusion in WS setting achieves perfect precision (100% compared to 75% for WoS) and higher recall (80% compared to 60%), suggesting that section titles help resolve ambiguities in identifying this rare class. Differently, for the *Future* class, in WS setting CiteFusion has increased precision (80% vs. 71.43%) but decreased recall (80% vs. 100%) w.r.t. WoS setting. This suggests that when section titles are included, the EC becomes more selective in assigning the *Future* label – i.e., makes fewer predictions for this class, but those predictions are more likely to be correct (higher precision). Here, the recall decline indicates that some *Future* instances are now missed.

Conversely, the *Motivation* class experiences a decline in precision (60% vs. 100% in WoS), despite maintaining identical recall (42.86%), suggesting that section titles may introduce noise for this class. This noise likely derives from the indirect interaction between section titles and the metaclassifier's base-models weighting mechanism. Indeed, *Motivation*-type citations appear in sections with ambiguous titles, such as “*Introduction*”, which lack a clear semantic association with motivational intent and are instead more closely related to other classes, such as *Background*. This missing association is particularly relevant when considering the severe imbalance of ACL-ARC, which do not provide the model with enough datapoints to infer it. Additionally, due to the scarcity of *Motivation*-type citations, the metaclassifier learns to give higher weight to positive probabilities from *Motivation*-based level-0 models. Therefore, in the WS setting, the presence of noisier signals derived from the ambiguity of the relation between section titles and class-specific cues, combined with the metaclassifier's tendency to overweight predictions from *Motivation*-specific models, results in the misclassification of these mixed signals as *Motivation*. This, in turn, leads to an increase in false positives.

Model	Background		Uses		ComOrCon		Extends		Motivation		Future	
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>
<i>CiteFusion (WS)</i>	81.48	92.96	78.26	69.23	85.71	72.00	100.0	80.00	60.00	42.86	80.00	80.00
<i>CiteFusion (WoS)</i>	76.47	91.55	75.00	69.23	81.25	52.00	75.00	60.00	100.0	42.86	71.43	100.0

Table 8. Class-specific Precision (*P*) and Recall (*R*) scores for the ACL-ARC Dataset.

Among majority classes in ACL-ARC, CiteFusion in WS setting has consistently higher performances w.r.t. WoS setting. For *Background*, precision increases from 76.47% to 81.48%, and recall improves slightly from 91.55% to 92.96%. Similarly, for *CompareOrContrast*, precision rises from 81.25% to 85.71%, and recall increases from 52% to 72%. Finally, for *Uses*, precision increases from 75.00% to 78.26%, and recall remains the same. These improvements likely derive from the stronger contextual alignment provided by section titles. These findings on class-specific precision and recall scores highlight that section titles disproportionately benefit rare classes associated with specific sections, such as *Extends*. However, their utility varies depending on the semantics of the class and the relevance of the titles. Overall, including section titles provide a benefit for most classes in ACL-ARC, as demonstrated by the increased overall performances in Macro-F1 and accuracy scores (**Table 6**).

Supervised Approaches – Comparison Between Heads

As demonstrated in **Table 3**, the FFNN meta-classifier consistently outperforms all other aggregation strategies in both accuracy and Macro-F1 scores across both datasets (SciCite and ACL-ARC) and settings (WS and WoS). The only exception is observed in the WoS setting for ACL-ARC, where Random Forests (RF) achieve marginally higher performance. A closer inspection of class-specific results reveals that RF excels in majority classes (e.g., *Background* and *CompareOrContrast*) but struggles with underrepresented intents. This behavior can likely be attributed to the inherent construction of RF. As a model designed to minimize the overall error rate, RF prioritizes identifying reliable predictions for majority classes, often at the expense of accuracy for minority classes (Chen et al., 2004). Consequently, while RF demonstrates strong performance in the WoS setting for the ACL-ARC dataset, its limited focus on underrepresented classes makes it less suitable as the final aggregator for CiteFusion. Considering this, we opted to retain the FFNN as the preferred meta-classifier, given its ability to balance performance across both majority and minority classes more effectively.

StackingC (Seewald, 2002), as described in *Section 3*, demonstrates robust performance in both datasets and settings, effectively leveraging class-specific binary predictions to mitigate data scarcity challenges. However, its lower performance relative to FFNN can be attributed to its inability to account for interdependencies between base model predictions across different classes. Unlike FFNN, which synthesizes cross-class interactions through a unified architecture, *StackingC* operates independently on each class, potentially limiting its capacity to resolve ambiguities arising from the overlapping of level-0 predictions of citation intents. This limitation is particularly evident in complex datasets like ACL-ARC, where the semantic boundaries between classes are less distinct.

The geometric framework employing Euclidean Distance to determine optimal weights for voting strategies, adapted from the paper of Wu and colleagues (2023) to our setting – as described in *Section 3* – yields performance improvements in SciCite compared to standard voting strategies, but produces mixed results in ACL-ARC. This discrepancy may stem from differences in the alignment of base model predictions between validation and test splits in ACL-ARC, probably due to the low number of instances, whereas SciCite exhibits greater consistency¹⁴. Such findings suggest that the geometric framework’s efficacy in finding optimal weights is related to the stability of component classifiers’ outputs, which is more easily achievable in datasets with a larger number of datapoints.

Overall, these results underscore the importance of selecting aggregation strategies tailored to dataset characteristics and level-0 classification processes. While FFNN’s adaptability makes it broadly effective, ensemble methods like *StackingC* and geometric frameworks offer complementary strengths in scenarios where class-specific or deterministic optimization is prioritized. Future work could explore hybrid approaches that integrate the geometric framework’s deterministic weighting with FFNN’s dynamic cross-class interactions to address these trade-offs.

4.2.1 Computational Instability: Observations on Robustness and Replicability

To further underscore the robustness of our strategy, beside its application to two different datasets in two different settings, we computed the average scores and their standard deviations across multiple runs of our classifiers in WS setting (see *Table 4*). For the SciCite dataset, mean accuracy and Macro-F1 scores are 90.53 ($\sigma = 0.14$) and 89.34 ($\sigma = 0.18$), respectively. These results not only surpass the previous SOTA model’s performances (*Table 5*) when averaged, but also demonstrate remarkable consistency, as our strategy surpasses SOTA in all ten experimental runs.

Similarly, for the ACL-ARC dataset, the mean accuracy and Macro-F1 scores are 80.44 ($\sigma = 0.44$) and 73.71 ($\sigma = 1.61$), respectively. Despite the relatively high standard deviation in Macro-F1, CiteFusion consistently outperformed SOTA models across all ten runs in both metrics (see *Tables 4 and 6*). The high standard deviation in Macro-F1 for ACL-ARC can be attributed to the extreme imbalance and variability among the 6 classes in the dataset, which poses challenges for classification tasks. Specifically, *Motivation* and *Future*, both underrepresented classes in ACL-ARC, exhibit greater fluctuations in SciBERT-based models’ performance across runs, contributing to the higher variance observed in Macro-F1 scores. Nevertheless, the consistent superiority of our strategy highlights its ability to effectively handle such complexities and deliver reliable performance even in challenging scenarios.

We collected intermediate training results of level-0 models across all ten runs for both datasets. The validation loss trends over these ten runs are visualized for each of the 6 level-0 models fine-tuned on SciCite in *Figure 3* and for each of the 12 level-0 models fine-tuned on ACL-ARC in *Figure 4*.

¹⁴ To clarify this point, it is important to note that the SciCite dataset contains a significantly larger number of instances in both the validation and test splits compared to ACL-ARC (see *Table 1*). This difference in dataset size has implications for the robustness of weight computation through the geometric framework. In the case of SciCite, the larger sample size ensures that even if outliers are present within the dataset, their impact on the class-specific weight calculation remains minimal due to the averaging effect across a greater number of instances. Conversely, in ACL-ARC, the relatively smaller number of instances makes the weight computation more sensitive to outliers. Even a small number of anomalous predictions from level-0 models can disproportionately influence the results, given the limited sample size.

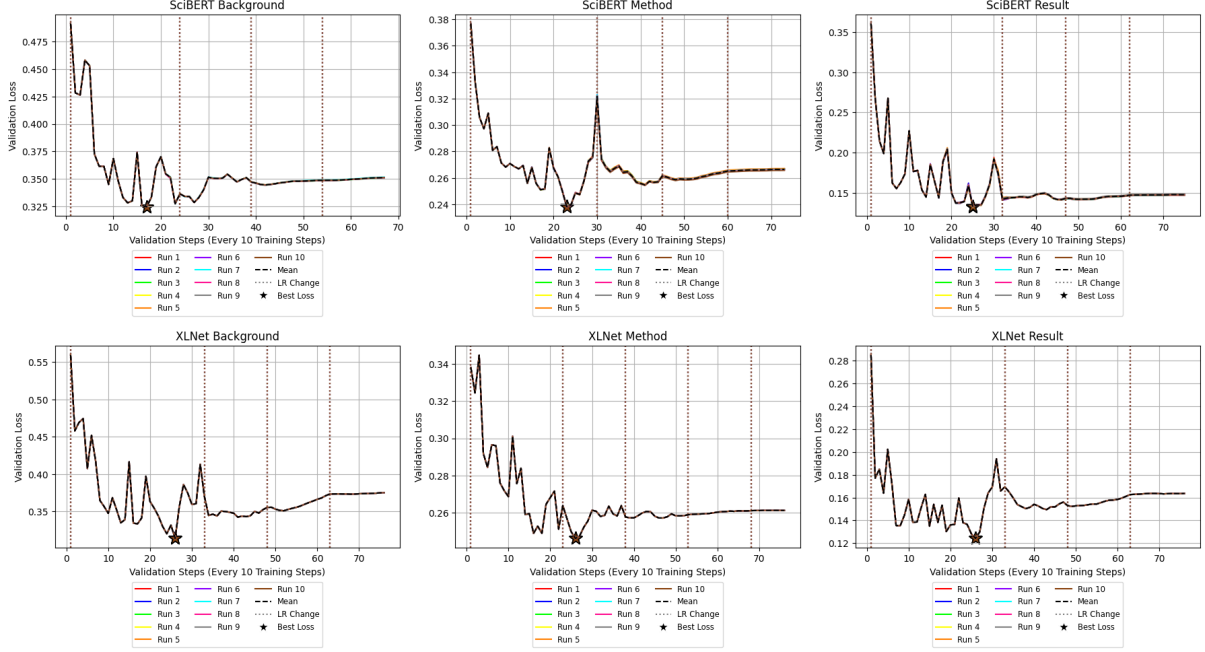


Figure 3. Visual comparison of validation losses across the fine-tuning steps for level-0 models trained on the SciCite dataset. The y-axis represents the validation loss scores, while the x-axis denotes the fine-grained evaluation steps. All models exhibited stable performance across the ten runs, as evidenced by the near-perfect overlap of the 11 lines in each plot (10 corresponding to individual models and one dashed line representing the mean). This overlap indicates consistent performance across validation steps. However, minor instability is observable in the SciBERT-based models after reaching the minimum validation loss. Upon closer inspection, slight separations between the lines at certain points suggest minimal variability in performance, though the overall stability remains completely unaffected since the best state of the models was yet saved when they reached their respective minimum.

Figure 3 illustrates that, across the repeated experimental runs on the SciCite dataset, all the ten base models within each class consistently converge to the same minimum in validation loss, with minimal variations recorded for SciBERT-based models (for full detailed reports of loss minima across runs, we refer the reader to **Appendix A.2**). Consequently, their performance remains stable across various applications of the entire ES, with a small standard deviation in final performance scores. According to the use of fine-grained evaluations, the final loaded state of these models is the one recorded at the evaluation step in which each of them reaches the minimum in validation loss. Thereby, the fluctuations in **Figure 3** are either negligible or minimal, resulting in equally robust ECs.

A similar trend is evident in **Figure 4** for most of the models trained on the ACL-ARC dataset, but with slight variations that account for the relatively higher standard deviation observed in this dataset. While the plots in **Figure 4** confirm that XLNet models exhibit perfect stability across different runs for different classes, they also reveal differences in training dynamics for certain classes. Minor discrepancies can be noted in SciBERT’s performance for the *Uses* and *Extends* classes; however, these do not significantly impact the overall performance of the ECs. Although the results indicate that the loss diverges at certain points, they also confirm that minima were consistently reached prior to such divergence. The variations in performance should be primarily attributed to SciBERT’s handling of the *CompareOrContrast*, *Motivation*, and *Future* classes, where higher differences in minima are observed across runs. Class-specific minima in validation loss for each run are provided in **Tables A.1 and A.2** of **Appendix A.2** to facilitate a detailed comparison with the visual trends depicted in **Figures 3 and 4**.

Our results and analyses of computational instability affirm the robustness and adaptability of our strategy across both datasets. Furthermore, the use of fine-grained evaluations and mixed precision demonstrates minimal impact on the performance, the robustness, and the replicability of CiteFusion in imbalanced settings. These findings also highlight the reliability of our binarized couple-based ensemble strategy and of fine-grained evaluations under conditions of severe imbalance and even with relatively low number of datapoints to tune level-0 models.

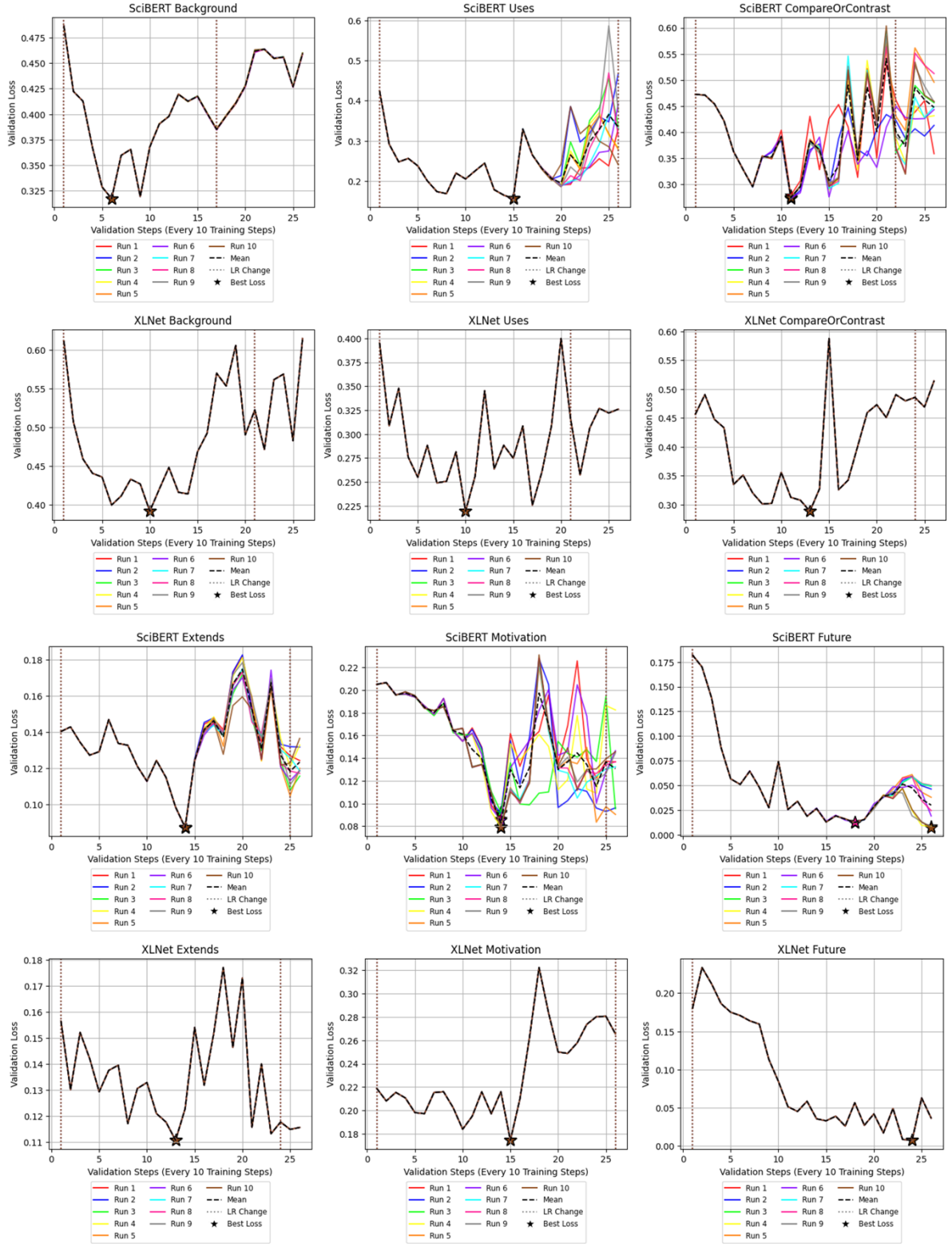


Figure 4. A visual comparison of validation losses across the fine-tuning steps for level-0 models trained on the ACL-ARC dataset reveals notable patterns. Specifically, in certain SciBERT-based models, variations in validation loss can be observed across different evaluations and experimental runs. For instance, the plots for SciBERT Motivation and SciBERT Future illustrate how the loss diverges during training in some instances, highlighting inconsistencies in convergence behavior across runs. These observations underscore the variability in training dynamics among specific model-class combinations.

4.3 Discussing the Classification Process of CiteFusion

Our analyses of SHAP results (reported in *Appendix A.1*) provide the top 15 tokens used by each model as features to positively classify an input sentence to their class. We can formally distinguish the two datasets in this level-0 discussion for a better understanding of the outcomes and to better assess the contributions of adding section titles within input sentences. Then, the discussion of the results related to the metaclassification process will be presented for both datasets together.

SciCite

Our analysis reveals that XLNet, with its standard vocabulary, tends to prioritize more general tokens, while SciBERT, leveraging the domain-specific training and its SciVOCAB (Beltagy et al., 2019), focuses on domain-specific tokens. This difference in vocabulary design results in SciBERT recognizing a broader set of features (over 11,000) that are closely aligned with scientific discourse, whereas XLNet identifies fewer and more generic features (approximately 9,700). This contrast is particularly pronounced when examining the most influential tokens identified by the two architectures within the same classes (see *Figures A.1 and A.2*). The implications of this divergence will become clearer in the subsequent section, where we analyze some misclassified citations.

Our analyses also demonstrate that models trained to classify sentences related to the *Background* class exhibit less informative features compared to those for *Method* and *Result*. For instance, top-contributing words for *Background* include terms like “circumference”, “ramification”, or “difficulty”, which lack a clear semantic attribution. In contrast, *Method* and *Result* classes are characterized by many semantically coherent terms such as “methodology” or “used” for the *Method* class, and “agrees” or “confirms” for *Result*, which align with their respective *semantic field* – a set of words that are related in meaning and cover a particular conceptual domain¹⁵ (Trier, 1931). This distinction is equally consistent across both WS and WoS settings. Furthermore, the mean SHAP values for *Background* are negative in all cases, whereas *Method* and *Result* exhibit positive means, particularly in WS experiments. This trend suggests that *Background* sentences contain less discriminative features on average, potentially due to higher variability in citation contexts and less intent-specific tokens.

The analysis of token contributions through SHAP values reveals distinct patterns across classes and model configurations. For the *Background* class, the sum of SHAP values is negative, indicating that models primarily identify features that exclude citations from this class rather than defining them as belonging to it, suggesting a classification more focused on “what is not”, thereby shaped by relevant features for other classes, negatively weighted by *Background*-specific models. A similar trend is observed for SciBERT-based models predicting *Method* and *Result* classes in the WoS setting. However, in the WS setting, these models better capture the defining characteristics of these classes, improving their ability to focus on what constitutes a *Method* or *Result*-related citation rather than relying on exclusionary cues. This shift, driven by the inclusion of section titles, is particularly prominent in SciBERT-based models, where SHAP values become more positive across all classes, including *Background*. In contrast, XLNet models exhibit a less pronounced change, likely due to their standard vocabulary, which does not emphasize the contextual significance of section titles as semantic features.

Another notable difference after including section titles is observed in the overall token contributions for the XLNet-Method model. Specifically, its raw score decreases from +32.12 in the WoS setting to +25.65 in the WS setting, reducing the importance assigned to influential tokens, which now receive lower scores. This highlights that the inclusion of section titles reshapes the perception of influential tokens for this model. Similar but less significant changes occur for XLNet-Background and XLNet-Result models. For instance, the XLNet-Background score shifts from -40.45 to -39.3, and the XLNet-Result score decreases from +11.71 to +10.96. These findings suggest that while the impact of section titles on token contributions varies across models, it is consistently present.

¹⁵ The intended conceptual domains are the ones described by the authors as descriptions of the labels used for their schemes in SciCite (Cohan et al., 2019) and ACL-ARC (Jurgens et al., 2018).

ACL-ARC

The analysis of SHAP values at level-0 for the ACL-ARC dataset reveals patterns that are mostly consistent with those observed in SciCite, while also showcasing distinct behaviors tied to the specific nature of different citation functions. Furthermore, our analysis shows results and feature patterns mostly consistent with the class-specific grammatical arguments – or features – identified by Jurgens et al. (2018).

In ACL-ARC, SciBERT-based models still exhibit domain specificity in the tokens they identify, though this is less pronounced and more subtle compared to SciCite. In the WS setting, SciBERT-based models tend to assign higher scores to intent-relevant tokens, a trend that becomes particularly evident when comparing common tokens between WS and WoS settings. For instance, the token “*Following*” for the *Uses* class receive a significantly higher mean score in the WS setting (+0.45) compared to WoS (+0.36), underscoring its importance in positively identifying citations with this intent. Similarly, the token “*limitation*” for the *CompareOrContrast* class increases from +0.03 in WoS to +0.14 in WS, highlighting its relevance. These tokens, which are intuitively linked to their respective citation intents in scientific articles, are not identified among the top 15 most positively attributed tokens by XLNet-based models. Instead, XLNet models tend to favor more general terms to strongly push their predictions towards a positive classification. Terms such as “*analogous*” or “*contrast*” guide the prediction of *CompareOrContrast*, while “*utilize*” or “*uses*” shape the *Uses* classification, further emphasizing XLNet models’ reliance on broader linguistic features rather than intent-specific cues.

Our analysis revealed that the *Background* class lacks a precise characterization of its intent also in ACL-ARC, consistently with prior findings on SciCite. In contrast, such characterization is clear for the remaining five classes. For instance, the *Uses* class in ACL-ARC is associated with terms similar to those defining the *Method* class in SciCite, such as “*utilize/s*” or “*applied*”, further supporting the alignment of these two intents with the same object property from CiTO (see **Table 2**). Similarly, the other classes exhibit distinct token-based characterizations. Examples include “*share*” or “*similar*” for the *CompareOrContrast* intent, “*elsewhere*” or “*earlier*” for the *Extends* class, “*inspired*” or “*motivated*” for *Motivation*, and “*promising*” or “*direction*” for the *Future* intent.

Our analysis of model-specific tokens reveals significant overlap with the grammatical features identified by Jurgens et al. (2018) for four classes in the ACL-ARC dataset. Specifically, the authors highlighted terms such as “*inspire/d*” for the *Motivation* class, which aligns with our findings in the top contributing features of our *Motivation*-specific models. Similarly, we observed comparable patterns for the *Uses* class with terms like “*use/s*” and “*follow/ing*”, as well as for *CompareOrContrast* with “*similar*”, and for the *Extends* intent with “*previous/ly*” and “*extend*”. Beyond these base grammatical features, our analysis extends to include additional terms that enrich the semantic characterization of each class. For instance, the *Motivation* class incorporates terms such as “*address*”, “*claim*”, and “*find*”, while the *Uses* class includes “*implementation*”, “*apply/ied*”, and “*employ*”. In the case of *CompareOrContrast*, we identified terms like “*advantage*”, “*outperform*”, and “*analogous*”, and for *Extends*, we found “*earlier*” and “*follows*”. However, for the *Background* and *Future* classes, no grammatical features were provided by Jurgens et al. (2018), rendering a direct comparison with our results unfeasible.

Finally, the analysis of the overall token contributions is consistent with the previously reported findings concerning the inclusion of section titles. For most classes, SciBERT models exhibit an improved ability to identify relevant positive features, likely attributable to their domain-specific training and vocabulary. Exceptions are observed for the *Extends* and *Future* classes, which are significantly underrepresented in the dataset. This underrepresentation may limit the model’s capacity to precisely characterize these classes, even after the incorporation of section titles, although it still assists in identifying what they do not represent, further supporting the decrease of the recall score for *Future* in WS setting when compared with WoS, as noted before. In ACL-ARC, the only notable difference in XLNet-based models following the inclusion of section titles is observed for XLNet-Background, which undergoes a negative shift of more than 8 points, suggesting a significant rebalance of the way in which this model perceive input sentences.

MetaClassification

The analysis of the FFNN metaclassifiers provided valuable insights into how predictions from base models are aggregated. In SciCite, across both WS and WoS settings, the metaclassifier effectively utilizes the correct level-0 models to identify classification intent. As shown in **Figures 5 and 6**, when classifying a citation into a specific class, the model positively relies on the level-0 PLMs trained for that class while assigning negative weights to high prediction values from the remaining four models.

The only exception is in classifying *Method* citations, where the metaclassifier tends to assign a relatively low positive importance to XLNet-Result. A plausible explanation is that the XLNet-Result model has identified certain textual patterns or signals that also characterize *Method*-type citations. As a result, when the metaclassifier encounters a high *Result*-class score from XLNet, it interprets this as weak but positive evidence for categorizing the citation as *Method*. Essentially, some linguistic cues used by the XLNet-Result model might overlap with the way in which *Method*-type citations are described. This misinterpretation is corrected with the inclusion of section titles, further demonstrating their utility. This is depicted in **Figure 6**, where the metaclassifier no longer misinterprets XLNet-Result predictions as pertaining to the *Method* class.

Furthermore, the dependence on XLNet-Method is less pronounced, likely due to the observed shift in token utilization following the inclusion of section titles. As previously noted, this model adjusted and balanced the classification process by generally reducing the scores assigned to tokens. Consequently, while its predictive power for binary classification diminished, it adopted a more cautious and conservative approach. The corrected interpretation of the outputs from XLNet-Result and the more cautious approach of XLNet-Method derived from the inclusion of section titles are also reflected in the class-specific precision and recall scores in **Table 7**, where the *Method* class is the only intent that sees an increase in both measures from WoS to WS setting.

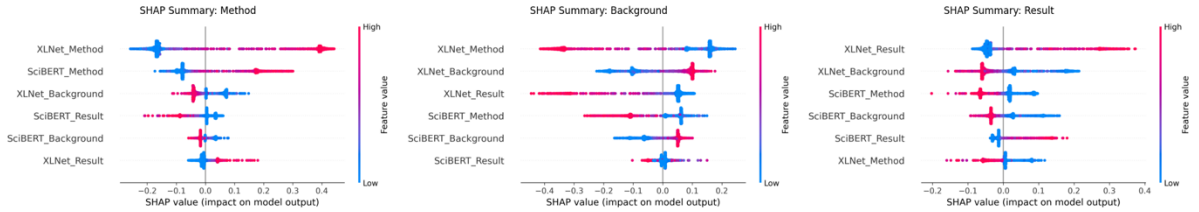


Figure 5. SHAP summary plots illustrating the contribution of each base model in classifying various intents within the SciCite dataset for the *WoS* setting.

An illustrative example of the aforementioned “*what is not*” kind of classifications can be observed in the XLNet-Method model when classifying instances of the *Background* class, in **Figure 5**. The plot reveals that the XLNet-Method model has high importance in this classification. Specifically, when the model predicts low percentage scores (represented by blue dots), it strongly contributes positively to classifying an instance as belonging to the *Background* class. Conversely, when the model predicts high scores (red dots), it plays a significant role in classifying the instance as not belonging to the *Background* class. This pattern underscores the metaclassifier’s sensitivity to XLNet-Method model’s predictions for distinguishing between the positive or negative outcome for the *Background* class.

The precision in identifying the appropriate models for classification is consistent across most classes also in ACL-ARC, with a few notable exceptions (see **Figure 7**). Beginning with the *Background* class, the FFNN metaclassifier misinterprets the role of XLNet-Background, assigning it negative importance when classifying instances of this class. Similarly, for the *CompareOrContrast* class, XLNet-Future, XLNet-Background, and XLNet-Extends are attributed positive importance, alongside the correctly identified models for this class. Notably, the contribution of SciBERT-CompareOrContrast is relatively minor compared to the other positively contributing features in this context. This may suggest some overlapping of contextual cues between the classes.

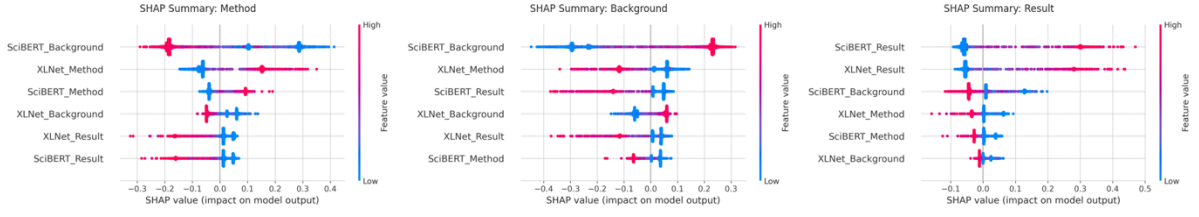


Figure 6. SHAP summary plots illustrating the contribution of each base model in classifying various intents within the SciCite dataset for the *WS* setting. As you can notice when comparing with Figure 5, XLNet-Result model is no more misinterpreted when it comes to classify Method class instances.

Further analysis of the plots in **Figure 7** reveals that for the most represented classes (*Background*, *Uses*, and *CompareOrContrast*), the model effectively leverages both positive and negative signals from various features, underscoring its reliance on diverse indicators for classification. However, this pattern does not hold for the most underrepresented classes (*Extends*, *Motivation*, and *Future*). For these classes, the metaclassifier outputs appears to depend almost exclusively on the models specifically designed for them, largely disregarding contributions from other features. This suggests a potential limitation in the model's ability to generalize or utilize broader contextual signals for less frequently occurring classes, which is probably derived from an overweighing of positive probabilities produced by the base models specifically tuned on these classes due to their underrepresentation.

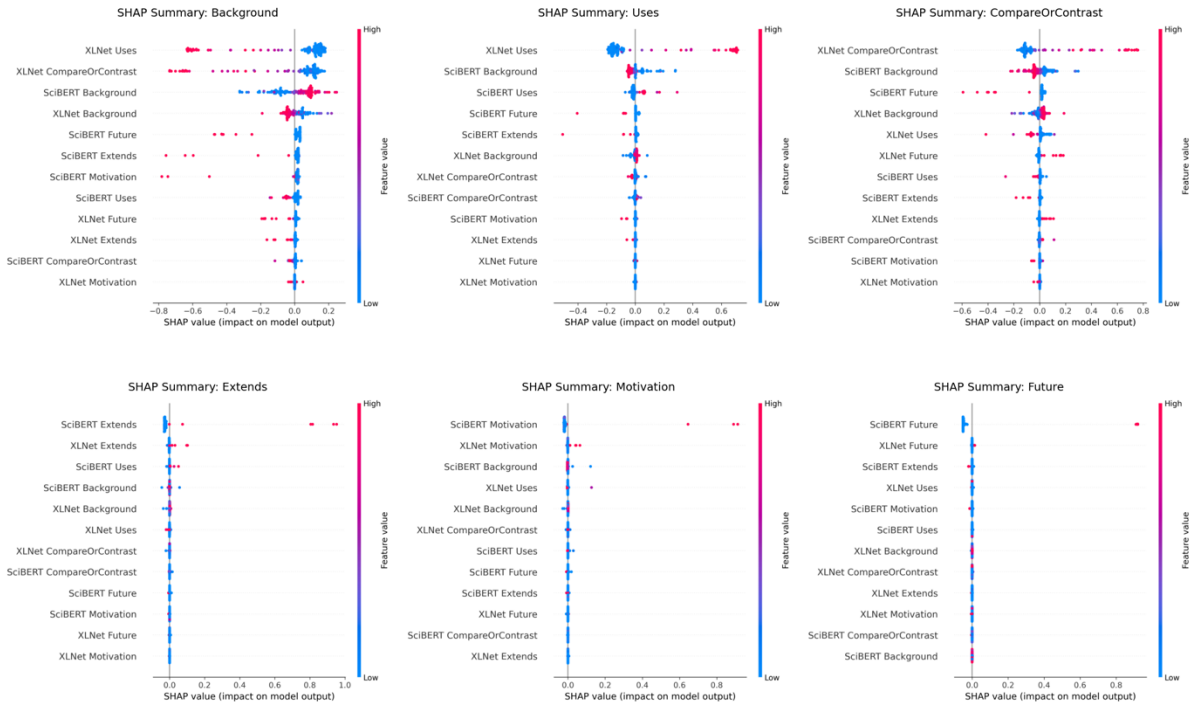


Figure 7. SHAP summary plots depicting the contribution of each base model in classifying various intents within the ACL-ARC dataset for the *WoS* setting.

The inclusion of section titles once again results beneficial for the overall predictions by mitigating the influence of individual token contributions and aligning the base models more closely with their respective correct classes. Upon inspecting the plots in **Figure 8**, can be noticed that the misinterpretation of the XLNet-Background feature as negatively contributing to the classification of *Background*-type instances is corrected after the addition of section titles. Similarly, the misinterpretations involving XLNet-Future, XLNet-Background, and XLNet-Extends for the *CompareOrContrast* class are resolved for the first two models and significantly mitigated for the third.

However, a challenge persists with XLNet-Extends in the *Extends* class, where it is still considered negatively related to the attribution of instances to this class. This issue was not present in the WoS setting but does not appear to interfere significantly with the final classification, as the class-specific scores for *Extends* still improve. Additionally, in the WoS setting, the classification process faced challenges due to the SciBERT-Uses model being positively associated with the *Extends* class, a problem that is now corrected in the WS setting by inversely aligning the SciBERT-Uses scores. However, a new issue arises with XLNet-Uses within the *Extends* class, where its predictions are positively aligned despite being expected to be the opposite. Fortunately, this discrepancy is relatively minor, as the SHAP values attributed to this feature remain small.

Two notable issues emerge instead for the *Motivation* and *Future* classes, where the metaclassifier's ability to correctly classify instances diminishes, as reflected in the performance scores presented in **Tables 6 and 8**. In these cases, although the metaclassifier appears to appropriately utilize relevant features and negatively weigh additional ones, its performance for these classes declines. This suggests that the more direct approach relying solely on pertinent features, as observed in the WoS setting, may have been advantageous for these two underrepresented intents.

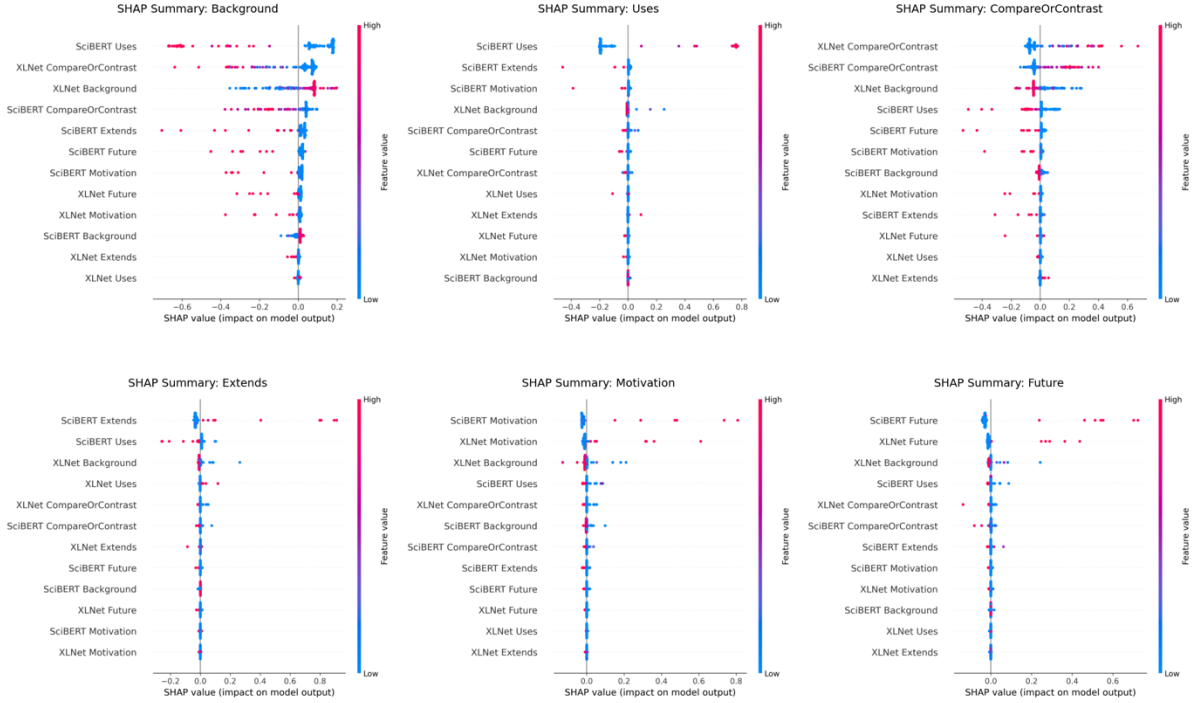


Figure 8. SHAP summary plots depicting the contribution of each base model in classifying various intents within the ACL-ARC dataset for the *WS* setting.

In summary, these additional findings enable us to assess the role of section titles before proceeding to examine some misclassified citations. Our analyses demonstrate that incorporating section titles into the input consistently enhances classification performance, improving both accuracy and macro-F1 across all experiments. When included in the input, section titles serve as contextual cues, with their impact being particularly pronounced in SciBERT-based models, which derive significant benefit from the extended context. Although XLNet also exhibits performance improvements, the effect is less pronounced, likely due to its vocabulary design (standard vs. domain-specific) and its general-domain training corpora (against the domain-specific pretraining of SciBERT). Overall, the WS setting outperforms the WoS setting across both datasets.

4.3.1 Local Interpretations of Misclassified Citations

SHAP values, as utilized thus far, have effectively demonstrated and provided support for elucidating the general classification dynamics of CiteFusion. This section, however, delves deeper into local explanations, offering contextual evidence regarding the types of errors made by our models. For simplicity, all experiments discussed in this section are conducted on the SciCite dataset in the WS setting. SciCite includes additional metadata related to the annotation process, specifically the confidence levels of annotators for certain sentences. To conduct an analysis of misclassified citations, we extracted all errors made by CiteFusion (WS)¹⁶ and focused on those instances accompanied by these metadata. This allowed us to evaluate the performance of our model in relation to both less certain annotations and those with complete agreement.

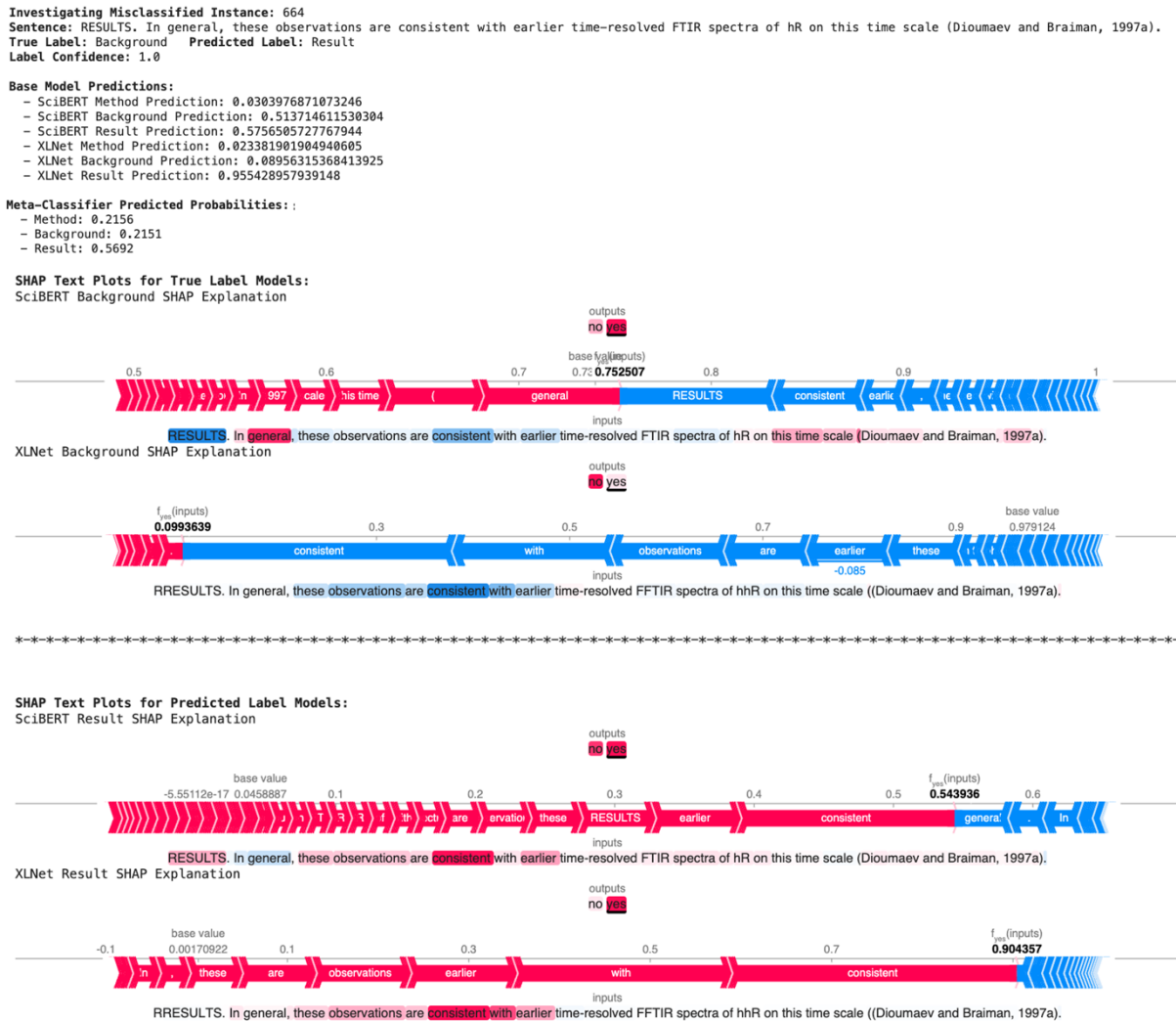


Figure 9. Analysis of the citation context depicted in the image, which was misclassified as Result by our model despite belonging to the Background class. The report begins by presenting the sentence to be classified, accompanied by the individual prediction scores from the base models and the final output scores from the FFNN meta-classifier. The four text plots are organized as follows: the first two plots illustrate the tokens with the highest SHAP values as attributed by the two base models specifically trained to identify the true label (in this case, the Background label). The remaining two plots represent how the models trained for the incorrectly predicted label (in this case, Result) classified the same instance. Red elements indicate positive contributions toward the positive binary classification decision ("yes"), while blue elements highlight negative contributions to it. These color-coded features reveal which parts of the input influenced the classification outcome for each respective model, and their width reveal how strong their influence was.

¹⁶ Out of the total 1,856 test instances, our model misclassified 179 instances.

The first sentence under examination is illustrated in **Figure 9**, where the element preceding the first period (“.”) represents the title of the section containing this citation context. The citation is classified as *Background* with a confidence score of 1 (on a scale from 0 to 1). The analysis presented in **Figure 9** elucidates how the base models within CiteFusion process the same citation text and assign scores at both the level-0 (binary classification) and level-1 (multi-class metaclassification) stages.

A key insight derived from the analysis in **Figure 9** is the different contribution of the same input sentence across various model architectures. As previously discussed, XLNet tends to emphasize more general tokens, whereas SciBERT assigns greater importance to domain-specific tokens. This distinction is particularly evident when examining the phrase “*FTIR spectra of hR on this time scale*”. SciBERT-based models attribute significant importance – either positive or negative, depending on the specific class – to this segment, likely because it relates to the scientific domain of the paper containing the citation, describing something somehow contained and characterized within SciVOCAB by SciBERT. In contrast, XLNet-based models do not assign any notable importance to this phrase.

Investigating Misclassified Instance: 692

Sentence: Discussion. In support of such a model, while the phenotype of Dusp6 mutant mice is distinct from that in zebrafish (see below), it is nevertheless incompletely penetrant [21], consistent with a more general role for dusp6 in maintaining a permissive range of ERK activity.

True Label: Background **Predicted Label:** Result
Label Confidence: 0.6065000295639038

Base Model Predictions:

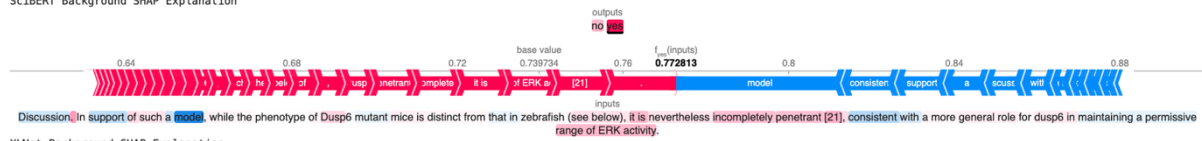
- SciBERT Method Prediction: 0.0203201025724411
- SciBERT Background Prediction: 0.27146080136299133
- SciBERT Result Prediction: 0.595274806022644
- XLNet Method Prediction: 0.0503542460501194
- XLNet Background Prediction: 0.27324894070625305
- XLNet Result Prediction: 0.6710962653160095

Meta-Classifier Predicted Probabilities:

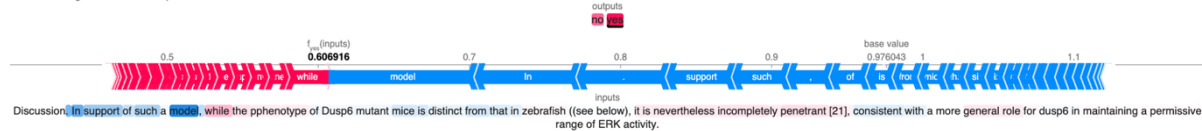
- Method: 0.2173
- Background: 0.2176
- Result: 0.5651

SHAP Text Plots for True Label Models:

SciBERT Background SHAP Explanation



XLNet Background SHAP Explanation



SHAP Text Plots for Predicted Label Models:

SciBERT Result SHAP Explanation



XLNet Result SHAP Explanation

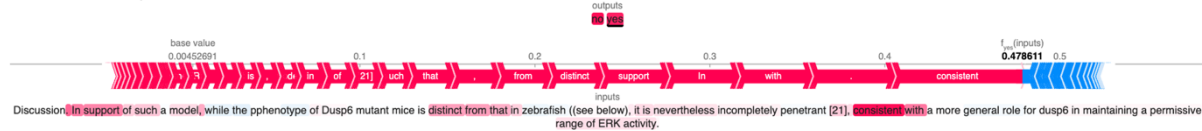


Figure 10. Analysis of the citation context depicted in the image, which was misclassified as Result by our model despite belonging to the Background class. Notably, the main difference with Figure 9 is the confidence score of the annotators for this citation context.

Similarly, elements of the citation such as author names or publication years are considered by SciBERT-based models to have a slight influence on the final prediction, while they are mostly disregarded by XLNet models. This pattern becomes even more pronounced when analyzing the role of the section title¹⁷. XLNet models completely overlook the section title in this case, whereas it plays a pivotal role in the predictions generated by SciBERT-based models. Specifically, for SciBERT models, the section title emerges as the most contributing token in classifying the sentence as not pertaining to the *Background* class, while ranks as the third most influential token in classifying it as belonging to the *Result* class. This underscores the critical role of structural elements when provided as contextual cues, such as section titles, in enhancing the performance of domain-specific models like SciBERT.

Furthermore, although the citation is labeled as a *Background*-type instance, we have reservations regarding the confidence score assigned by the annotators. The sentence depicted in **Figure 9** seems to demonstrate a greater alignment with the class identified by our CiteFusion model.

Another example of a misclassified citation is presented in **Figure 10**, this time annotated with a confidence score of 0.61. Our model incorrectly predicts the citation as belonging to the *Result* class, whereas it is labeled as a *Background* citation. Consistent with previous observations, SciBERT demonstrates alignment with the scientific vocabulary and domain-specific terminology, while XLNet models continue to disregard section titles, instead assigning higher scores to more general elements.

Additional misclassified citations include:

- “Discussion. Several studies have now shown that DNA methylation changes with age at different genomic locations, the direction, and rate of change [9, 14, 33].”, misclassified as *Background* while labeled as *Method* with a confidence of 0.76;
- “2. Methods. Finally, the data regarding the obstetric history of the subjects were retrieved from the Danish Medical Birth Registry (MBR) which includes data on all live births, stillbirths and infant deaths in Denmark (Knudsen and Olsen, 1998).”, misclassified as *Method* while labeled as *Background* with a confidence score of 1;
- “DISCUSSION. The rapid disappearance of cyclin mRNA after RNAi treatment is consistent with the currently-held notion that dsRNA targets the gene-specific destruction of mRNA in a wide variety of eukaryotic cells (Fire et al., 1998; Tabara et al., 1998; Tabara et al., 1999; Sanchez-Alvorado and Newmark, 1999; Grishok et al., 2000; Boscher and Labouesse, 2000; Klink and Wolniak, 2000).”, labeled as *Background* with 0.76 confidence, but classified by CiteFusion as *Result*.

While our model occasionally misclassifies certain instances due to mathematical elements and formulas contained in it, or due to ambiguities arising from sentence structures, it is important to note that the reliability of the annotations themselves may also be a contributing factor. This observation holds true even for instances where the confidence scores associated with the labels are relatively high. Such findings underscore the need for continuous refinement and improvement in the quality of datasets used within the domain of Citation Intent Classification. Furthermore, they highlight the absence of a definitive gold standard that can serve as a robust foundation for training and developing highly reliable models. Establishing such a standard remains an open challenge and a critical area for future research, emphasizing the importance of collaborative efforts to enhance dataset accuracy and consistency in this field.

¹⁷ In **Figure 9**, the input sentence reported under the XLNet models contains certain errors. It is important to note that these discrepancies are not related to the classification process or the computation of SHAP values. Instead, they arise during the detokenization of the input performed by SHAP to generate the text plot. Consequently, these errors have no bearing on the classification outcomes or the associated scores.

4.3.2 Concluding Remarks on Explainers and their Role

Although the primary focus of this research was not to explain the decision-making processes of the developed models, and despite SHAP having certain limitations – particularly its inability to fully account for feature dependencies in transformer-based models (Gohel et al., 2021) – we have nonetheless provided some clarification regarding how the models generated in this study classify citation contexts. These analyses shed light on the respective strengths of the two PLM architectures, as well as on the aggregation process facilitated by the FFNN, which effectively integrates the contributions of all the level-0 models, differently from other strategies more focused on the specificity of each class, such as *StackingC* (Seewald, 2002), or the *ED+MLR adaptation of the Geometric Framework* (Wu et al., 2023), substantially based on voting strategies.

Overall, at level-0, SciBERT and XLNet demonstrate complementary strengths: SciBERT excels in identifying domain-specific tokens, while XLNet focuses on more general terms. At level-1, the metaclassifier synthesizes these specialized predictions by upweighting the base models that align with the correct class and downweighting predictions from non-matching base models in most cases. This synergy contributes to robust performance across different classes, efficiently mitigating the influence of class imbalance in both datasets.

Finally, to the best of our knowledge, no prior studies have systematically explored the semantic features that characterize citation functions based on the intent categories described by SciCite and ACL-ARC schemes. Furthermore, we extended from the explanations of raw citation contexts by assessing the role of structural features (i.e., section titles) when provided as contextual elements. Section titles are analyzed when considered by PLMs as semantic and grammatical components of citations, and their presence demonstrated to generally amplify the ability of base models to retrieve the constituting component of what class-specific citations are, also for underrepresented classes, thereby reducing models’ reliance on exclusionary cues. Although this investigation falls outside the primary scope of our work, we ultimately provided model-dependent generalizations regarding the types of words, or tokens, that influence the understanding and perception of different citation intents across two different PLM architectures. These insights lay foundational groundwork for future research in this domain.

5 CIC Application

This chapter outlines the implementation of a web application designed to deploy the state-of-the-art (SOTA) ensemble model developed in this study, specifically for the SciCite dataset. The decision to focus exclusively on CiteFusion models trained using SciCite is driven by the higher reliability observed on this dataset compared to ACL-ARC. Additionally, we find the classification schema employed by SciCite to be more robust and better suited for the task, further justifying its selection for this deployment.

The application is built using *Flask*¹⁸, a lightweight and flexible web framework, which enables the seamless integration of various model variants, including those trained in WoS (without section titles) setting, into a user-friendly web-based platform, currently available¹⁹. Additionally, this chapter provides a general overview of the application's development process and its functional capabilities.

5.1 General Aim and Description

The application was developed within the *OpenCitations* infrastructure²⁰ as part of the *GraspOS* project²¹. This initiative aims to establish a robust data infrastructure and promote an ethical research assessment system grounded in Open Science (OS) principles across Europe. GraspOS contributes to the *European Open Science Cloud* (EOSC) ecosystem by integrating tools that monitor research service usage and advocate for the adoption of OS principles. Within this context, citation data plays a pivotal role in fostering openness, legitimacy, and knowledge sharing among academic communities, aligning closely with the objectives of both GraspOS and the broader Open Science movement.

The classifier presented in this study forms part of a larger application designed to automate the extraction and classification of citation intents from PDF files of scholarly works. Specifically, the classifier component, which was developed as part of this research and detailed in Paolini (2024a), encompasses a backend system capable of loading the various models generated in this work, preprocessing input data into a compatible format, and ultimately classifying citation contexts. The tool was implemented using Flask, Python, HTML, CSS, and JavaScript, and is currently accessible to the public in its Beta version. This development represents a significant step toward enhancing the automation and accuracy of citation intent classification within academic research workflows.

5.2 Design and Implementation

The central component of the software is the *Predictor* object, which manages prediction tasks, allocates GPU resources when available, and performs tokenization processes specifically tailored for both section-based and non-section-based data. Additionally, it generates a downloadable JSON file containing the classification results.

The backend architecture comprises several key components:

- *EnsembleClassifier*: This module loads base models from the server in accordance with the instructions provided by the *Predictor*.
- *DataProcessor*: Responsible for data preprocessing tasks, including reading, formatting, and performing structural integrity checks on the input data.
- *MetaClassifierSection* and *MetaClassifierNoSection*: These modules define the metaclassifier architectures designed to handle the two distinct data scenarios – section-based and non-section-based inputs.

The backend system processes the input data and employs ensemble models to classify citation contexts based on the selected operational mode. To ensure the reliability of classifications, a human-defined threshold is integrated

¹⁸ <https://flask.palletsprojects.com/en/3.0.x/>

¹⁹ <http://137.204.64.4:81/cic/>

²⁰ <https://opencitations.net/>

²¹ <https://www.graspos.eu/>

into the system. This threshold evaluates the output probabilities generated by the classification process. If none of the three classes achieves a confidence score exceeding 90%, the classification is deemed unreliable. In such cases, the result is mapped to <http://purl.org/spar/cito/citesForInformation>, a general-purpose object property within the Citation Typing Ontology (CiTO) used to represent non-characterized citations. This mapping enhances interoperability while addressing cases where classification confidence is insufficient.

5.3 User Interface

The interface developed using Flask, HTML, CSS, and JavaScript, adopts a minimalistic design while ensuring sufficient user-friendliness in its initial release (see **Figure 11**). It enables users to input data either as a list of text tuples, or as JSON files, to then select from the following analysis modes:

- *Mixed Mode*: This mode utilizes both ensembles to dynamically identify citations with and without section titles and applies the corresponding model for classification.
- *With Section Titles*: Designed for sentences that include section titles, this mode employs the ensemble trained on data containing section titles.
- *Without Section Titles*: Intended for raw citation contexts, this mode uses the ensemble trained on pure citations without section titles.

Select your classification mode:

Mixed With Section Titles Without Section Titles

Scegli file nessun file selezionato Classify JSON

```
[
  ("Introduction", "In his 1945 essay 'As We May Think', Vannevar Bush observed how 'publication has been extended far beyond our present ability to make real use of the record' [Bush, 1945]."),
  ("Introduction", "Licklider expanded on this with the vision of a symbiotic relationship between humans and machines. Computers would take care of routine tasks such as storage and retrieval, 'preparing the way for insights and decisions in scientific thinking' [Licklider, 1960]."),
  ("Introduction", "Computing has indeed revolutionized how research is conducted, but information overload remains an overwhelming problem [Bormmann and Mutz, 2014]."),
  ("Introduction", "In May 2022, an average of 516 papers per day were submitted to arXiv [arXiv, 2022]. Beyond papers, scientific data is also growing much more quickly than our ability to process it [Marx, 2013]. As of August 2022, the NCBI GenBank contained 1.49 × 1012 nucleotide bases [GenBank, 2022].")
]
```

Classify

Figure 11. User input interface. The user can decide the classification mode, and whether to upload data in text (with a list of tuples containing section titles – if possible – and citation contexts) or JSON format.

Sentence: 0

SECTION: Introduction

CITATION: In his 1945 essay 'As We May Think', Vannevar Bush observed how 'publication has been extended far beyond our present ability to make real use of the record' [Bush, 1945].

SCIBERT MET POSITIVE PROBABILITY: 0.027594441547989845

SCIBERT BKG POSITIVE PROBABILITY: 0.9479248523712158

SCIBERT RES POSITIVE PROBABILITY: 0.007289402186870575

XLNET MET POSITIVE PROBABILITY: 0.021375877782702446

XLNET BKG POSITIVE PROBABILITY: 0.9609192609786987

XLNET RES POSITIVE PROBABILITY: 0.0028760619461536407

MET ENSEMBLE CONFIDENCE: 0.00039628162630833685

BKG ENSEMBLE CONFIDENCE: 0.9989715814590454

RES ENSEMBLE CONFIDENCE: 0.0006320800166577101

FINAL PREDICTION: obtainsBackgroundFrom (BACKGROUND)

Figure 12. Visualization of the classification results for a single sentence.

After the backend loads the models and processes the input citation sentences, the classification results are displayed. Upon completion of the analysis, the application provides detailed results for all predictions. These results include confidence scores from the level-0 models, the metaclassifier's confidence, and the final classification mapped to CiTO (**Figure 12**). Furthermore, the tool offers an option to download comprehensive JSON files containing both the citation data and the generated classification metadata.

6 Conclusion and Future Works

Understanding the motivations behind academic citations is crucial for analyzing scholarly discourse. Citation Intent Classification (CIC) aims to provide deeper insights into the underlying reasons for citations, thereby enhancing research evaluation and improving the transparency and reliability of academic communication. This study advances this objective by introducing advanced ensemble models for the CIC task, which demonstrate proficiency in addressing dataset imbalance-related challenges. As illustrated in *Tables 5 and 6*, the ensemble models developed in this research outperform the previous state-of-the-art (SOTA) results for CIC on both the SciCite and ACL-ARC benchmarks. These findings underscore the critical role of the stacked and binary-coupled architecture employed in this study. Furthermore, we observed that the results remain robust and consistently outperform SOTA across all the 10 runs of our experiment for both SciCite and ACL-ARC in WS (with section titles) setting, even when mixed-precision training was utilized together with fine-grained evaluations.

To further elucidate the decision-making processes of the ensemble models and provide transparency into their predictions, we employed SHAP (SHapley Additive exPlanations) for interpreting both level-0 binary predictions and the FFNNs’ multi-class metaclassifications, and mapped the original class labels to standard object properties from the Citation Typing Ontology (CiTO). SHAP-based analyses revealed significant differences in how SciBERT and XLNet prioritize and assign importance to tokens when classifying citation contexts, highlighting the role of domain-specific training and vocabularies.

These differences are particularly evident when examining the SHAP values assigned to the most important tokens for each model, both in a general and local way. SciBERT’s ability to recognize a broader range of scientific terminology results in more distributed scores, whereas XLNet’s focus on general terms leads to higher confidence in standard cues. The use of SHAP not only highlights the unique strengths of each architecture, but also provides insights into their complementary roles within the ensemble framework of CiteFusion, by explaining the way in which the FFNNs trust and merges different predictions from the base models. By leveraging SHAP, we were able to validate the robustness of our ensemble architecture and demonstrate how the integration of these two models enhances overall classification performance. Furthermore, this interpretability aids in understanding the factors contributing to misclassifications, thereby guiding future refinements to the system.

In addition to these contributions, we developed and publicly released a preliminary version of a web-based application designed to automatically classify citation contexts using CiteFusion models trained on SciCite in both WS and WoS settings. Building upon the original SciCite schema, we extended the framework by mapping the three initial labels to CiTO object properties to enhance interoperability. Additionally, we introduced a more general citation function to classify citation contexts for which our model lacks sufficient confidence.

This concluding section provides a more detailed analysis of the role of section titles in the structural composition of sentences intended for classification. It also outlines potential future work aimed at improving the reliability of the web-based application. Finally, it offers a comprehensive summary of the research contributions presented in this study.

6.1 Structural Features: Future Directions

Our experimental evaluations on both SciCite and ACL-ARC datasets demonstrate that incorporating section titles as part of the input text significantly enhances the performance of citation intent classification. Across all experimental configurations, the WS setting consistently surpasses the WoS baseline. Including section titles within citation contexts provide for additional semantical cues that help in reshaping – and improving – the way in which level-0 models perceive input sentences. Overall, section titles provide critical contextual information that enhances the models’ ability to discern fine-grained distinctions among citation types, especially in cases of imbalanced data. This contextual enrichment contributes to the robust and state-of-the-art performances achieved by CiteFusion.

The importance of section titles within the sentences to be classified suggests several avenues for enhancing performance in the CIC domain and achieving more reliable systems. One promising direction involves integrating PLMs and ensemble strategies into the broader framework of *Neuro-Symbolic Systems* (Yu et al., 2023). Such frameworks enable classifications that are not solely reliant on the intrinsic semantics of textual contexts but also incorporate structured data sources, such as Knowledge Graphs (KGs), which facilitate logical reasoning operations (Garcez et al., 2015; Daniele et al., 2022). Knowledge Graphs constructed from citation data can encapsulate logical rules and relationships, which may prove to be highly informative for the CIC task. Given our findings on the effectiveness of integrating structural elements like section titles, leveraging both the semantic and structural significance of these features within a broader context may theoretically yield further performance improvements.

Thereby, the generation of Knowledge Graph Embeddings (KGEs) offers a practical means to integrate structured data with embeddings derived from PLMs. This integration can be achieved through various methodologies, although a detailed exploration of these techniques falls outside the scope of this study. Nevertheless, combining logical rules, symbolic representations, and sub-symbolic elements with the intrinsic semantics of citation contexts extracted via LMs should theoretically provide a more structured and interpretable approach to CIC.

Enhanced explainability of model predictions, as emphasized by Pan et al. (2024), is crucial for developing robust and reliable systems suitable for deployment in production environments. By bridging the gap between semantic understanding and structural/logical reasoning, such an approach holds significant potential for advancing the state-of-the-art in CIC and fostering greater transparency in automated citation analysis systems.

6.2 Citation Intent Classifier: Future Releases

The web-based application is currently in its initial release and has undergone significant improvements since its preliminary version, which was developed as part of the master’s thesis titled “*Enhanced Citation Intent Classification with Population-based Training, Ensemble Strategies, and Language Models*” (Paolini, 2024c). Building on the work presented thus far, we aim to enhance the tool by incorporating explanations for classified sentences using SHAP, which provides APIs that enable dynamic and engaging visualization of results, such as the ones shown in **Figures 9 and 10**. Furthermore, users will be provided with the possibility to manually set different thresholds to distinguish between reliable and unreliable classifications. By integrating these explanations with user-defined thresholds tailored to specific needs, we aim to create a more robust, transparent, and user-friendly tool that can potentially be adapted to a variety of use cases.

In addition to these improvements, we plan to further enhance the performance of the backend models by leveraging an expanded dataset and incorporating KGEs into the classification pipeline. This integration is expected to improve the system’s ability to capture structured relationships within citation data, thereby enhancing classification accuracy and interpretability.

6.3 Final Remarks

In summary, this research positions the field of Citation Intent Classification at a critical juncture between the need to deepen the understanding of academic discourse and the advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI). A central contribution of this study is the emphasis on the necessity of a comprehensive and meticulously curated dataset for the CIC task, supported by empirical evidence demonstrating the pivotal role that data quantity and quality play in determining outcomes. For instance, the low performance observed across all models for the *Background* classes in both SciCite and ACL-ARC in terms of comprehensibility highlights this dependency. Similarly, the challenges associated with the underrepresented classes (*Result* in SciCite, and *Extends*, *Motivation*, and *Future* in ACL-ARC) underscore the limitations posed by an insufficient number of data points for developing robust models. This is evident not only in the class-specific performance metrics for these intents, but also in the aggregated explanations generated using SHAP.

Despite these challenges, the key contribution of this work is the development of *CiteFusion*, an ensemble strategy that achieves state-of-the-art performance for the Citation Intent Classification task on both the SciCite and ACL-ARC datasets. To ensure transparency and reproducibility, we publicly release all code (Paolini, 2024d) and models developed during this study, including versions tailored for both with section titles (WS) and without section titles (WoS) settings:

- CiteFusion_{SciCite} (WS) (Paolini, 2024b),
- CiteFusion_{SciCite} (WoS) (Paolini, 2024a),
- CiteFusion_{ACL-ARC} (WS) (Paolini, 2025b),
- CiteFusion_{ACL-ARC} (WoS) (Paolini, 2025a).

This research establishes a foundation for advancing the Citation Intent Classification task and guiding the development of future datasets. These efforts should be closely aligned with the evolving field of Explainable AI (XAI), as this study demonstrates that XAI not only enhances the interpretability of model behavior but also plays a crucial role in identifying and mitigating potential biases within the developed tools. Additionally, we provide a clearer – even if model-dependent – understanding of citation intents by highlighting the tokens and words that different models consider relevant. This analysis enables a preliminary semantic characterization of the intents used in SciCite and ACL-ARC, while also revealing significant overlaps between classes (e.g., *Background* with *Background*, and *Method* with *Uses*) across datasets. These overlaps are further enriched through annotation with CiTO object properties, facilitating their intersection and comparative analysis. In conclusion, by contributing to the field of Citation Intent Classification, this work fosters interdisciplinary dialogue and lays the groundwork for future innovations and investigations in the domain.

Acknowledgments

This work has been partially funded by the European Union's Horizon Europe framework programme under Grant Agreements No 101095129 (*GraspOS Project*) and No 101188018 (*GRAPHIA Project*).

Bibliography

- Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2010*, 11–15.
- Aldhyani, T. H. H., & Alkahtani, H. (2021). A Bidirectional Long Short-Term Memory Model Algorithm for Predicting COVID-19 in Gulf Countries. *Life*, 11(11), 1118. <https://doi.org/10.3390/life11111118>
- Artur D 'Avila Garcez, Besold, T. R., Raedt, L. D., Földiák, P., Hitzler, P., Icard, T., Kai-Uwe Kühnberger, Lamb, L. C., Mikkilainen, R., & Silver, D. L. (2015). *Neural-Symbolic Learning and Reasoning: Contributions and Challenges*. <https://doi.org/10.13140/2.1.1779.4243>
- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1607.06450>
- Bakhti, K., Niu, Z., & Nyamawe, A. S. (2018). Semi-Automatic Annotation for Citation Function Classification. *2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (IC-CAIRO)*, 43–47. <https://doi.org/10.1109/ICCAIRO.2018.00016>
- Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. <https://doi.org/10.48550/ARXIV.1903.10676>
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166. <https://doi.org/10.1109/72.279181>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/BF00058655>
- Chen C., Liaw A., & Breiman L. (2004). *Using Random Forest to Learn Imbalanced Data*.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. <https://doi.org/10.3115/v1/D14-1179>
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.1412.3555>
- Ciccarese, P., Shotton, D., Peroni, S., & Clark, T. (2014). CiTO + SWAN: The web semantics of bibliographic records, citations, evidence and discourse relationships. *Semantic Web*, 5(4), 295–311. <https://doi.org/10.3233/SW-130098>
- Cohan, A., Ammar, W., Van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of the 2019 Conference of the North*, 3586–3596. <https://doi.org/10.18653/v1/N19-1361>
- Cohan, A., & Goharian, N. (2015). Scientific Article Summarization Using Citation-Context and Article's Discourse Structure. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 390–400. <https://doi.org/10.18653/v1/D15-1045>
- Cornegruta, S., Bakewell, R., Withey, S., & Montana, G. (2016). Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, 17–27. <https://doi.org/10.18653/v1/W16-6103>
- Daniele, A., Campari, T., Malhotra, S., & Serafini, L. (2022). *Deep Symbolic Learning: Discovering Symbols and Rules from Perceptions*. <https://doi.org/10.48550/ARXIV.2208.11561>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- Dong, C., & Schafer, U. (2011). Ensemble-style Self-training on Citation Classification. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 623–631.
- Dutta, A., Kumar, S., & Basu, M. (2020). A Gated Recurrent Unit Approach to Bitcoin Price Prediction. *Journal of Risk and Financial Management*, 13(2), 23. <https://doi.org/10.3390/jrfm13020023>
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, 148–156.
- Garfield, E. (1964). Can Citation Indexing Be Automated? *Can Citation Indexing Be Automated?*

- Garzone, M., & Mercer, R. E. (2000). Towards an Automated Citation Classifier. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence* (Vol. 1822, pp. 337–346). Springer Berlin Heidelberg.
https://doi.org/10.1007/3-540-45486-1_28
- Gohel, P., Singh, P., & Mohanty, M. (2021). *Explainable AI: Current status and future directions* (arXiv:2107.07045). arXiv. <http://arxiv.org/abs/2107.07045>
- Graves, A. (2013). *Generating Sequences With Recurrent Neural Networks* (Version 5). arXiv.
<https://doi.org/10.48550/ARXIV.1308.0850>
- Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying Important Citations Using Contextual Information from Full Text. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–8.
<https://doi.org/10.1109/JCDL.2017.7991558>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Huang, Y., Feng, X., Li, B., Xiang, Y., Wang, H., Qin, B., & Liu, T. (2024). *Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration* (arXiv:2404.12715). arXiv.
<http://arxiv.org/abs/2404.12715>
- Jiang, D., Ren, X., & Lin, B. Y. (2023). *LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion* (arXiv:2306.02561). arXiv. <http://arxiv.org/abs/2306.02561>
- Jochim, C., & Schutze, H. (2012). Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. *International Conference on Computational Linguistics*.
- Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, 6, 391–406. https://doi.org/10.1162/tacl_a_00028
- Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, 244, 122778. <https://doi.org/10.1016/j.eswa.2023.122778>
- Kim, H.-C., Pang, S., Je, H.-M., Kim, D., & Yang Bang, S. (2003). Constructing support vector machine ensemble. *Pattern Recognition*, 36(12), 2757–2767. [https://doi.org/10.1016/S0031-3203\(03\)00175-4](https://doi.org/10.1016/S0031-3203(03)00175-4)
- Kunnath, S. N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. *Quantitative Science Studies*, 2(4), 1170–1215. https://doi.org/10.1162/qss_a_00159
- Lahiri, A., Sanyal, D. K., & Mukherjee, I. (2023). CitePrompt: Using Prompts to Identify Citation Intent in Scientific Papers. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 51–55.
<https://doi.org/10.1109/JCDL57899.2023.00017>
- Latif-Shabgahi, G. R. (2004). A novel algorithm for weighted average voting used in fault tolerant computing systems. *Microprocessors and Microsystems*, 28(7), 357–361.
<https://doi.org/10.1016/j.micpro.2004.02.006>
- Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139, 107539. <https://doi.org/10.1016/j.chb.2022.107539>
- Li, Z., & Ho, Y.-S. (2008). Use of citation per publication as an indicator to evaluate contingent valuation research. *Scientometrics*, 75(1), 97–110. <https://doi.org/10.1007/s11192-007-1838-1>
- Liu, L., Wu, X., Li, S., Li, Y., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection. *BMC Medical Informatics and Decision Making*, 22(1), 82. <https://doi.org/10.1186/s12911-022-01821-w>
- Liu, S., Wang, Y., Zhang, J., Chen, C., & Xiang, Y. (2017). Addressing the class imbalance problem in Twitter spam detection using ensemble learning. *Computers & Security*, 69, 35–49.
<https://doi.org/10.1016/j.cose.2016.12.004>
- Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malignieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106, 102301.
<https://doi.org/10.1016/j.inffus.2024.102301>
- Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization* (Version 3). arXiv.

<https://doi.org/10.48550/ARXIV.1711.05101>

- Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (arXiv:1705.07874). arXiv. <http://arxiv.org/abs/1705.07874>
- Mercier, D., Rizvi, S., Rajashekar, V., Dengel, A., & Ahmed, S. (2021). ImpactCite: An XLNet-based Solution Enabling Qualitative Citation Impact Analysis Utilizing Sentiment and Intent: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, 159–168. <https://doi.org/10.5220/0010235201590168>
- Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2), 757–774. <https://doi.org/10.1016/j.jksuci.2023.01.014>
- Monteiro, J. P., Ramos, D., Carneiro, D., Duarte, F., Fernandes, J. M., & Novais, P. (2021). Meta-learning and the new challenges of machine learning. *International Journal of Intelligent Systems*, 36(11), 6240–6272. <https://doi.org/10.1002/int.22549>
- Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving Predictions using Ensemble Bayesian Model Averaging. *Political Analysis*, 20(3), 271–291. <https://doi.org/10.1093/pan/mps002>
- Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, 5(1), 86–92. <https://doi.org/10.1177/030631277500500106>
- Nanba, H., Kando, N., & Okumura, M. (2011). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1), 117–134. <https://doi.org/10.7152/acro.v11i1.12774>
- Nazir, S., Asif, M., Ahmad, S., Bukhari, F., Afzal, M. T., & Aljuaid, H. (2020). Important citation identification by exploiting content and section-wise in-text citation count. *PLOS ONE*, 15(3), e0228885. <https://doi.org/10.1371/journal.pone.0228885>
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580–3599. <https://doi.org/10.1109/TKDE.2024.3352100>
- Paolini, L. (2024a). *CiteFusion WoS SciCite*. Zenodo. <https://doi.org/10.5281/ZENODO.14989091>
- Paolini, L. (2024b). *CiteFusion WS SciCite*. Zenodo. <https://doi.org/10.5281/ZENODO.14989192>
- Paolini, L. (2024c). *Enhanced Citation Intent Classification with Population-Based Training, Ensemble Strategies, and Language Models*. <https://doi.org/10.5281/ZENODO.11535143>
- Paolini, L. (2024d). *CiteFusion Experimental Notebooks* [Computer software]. Zenodo. <https://doi.org/10.5281/ZENODO.14988872>
- Paolini, L. (2025a). *CiteFusion WoS ACL-ARC*. Zenodo. <https://doi.org/10.5281/ZENODO.14989462>
- Paolini, L. (2025b). *CiteFusion WS ACL-ARC*. Zenodo. <https://doi.org/10.5281/ZENODO.14989775>
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). *On the difficulty of training Recurrent Neural Networks* (arXiv:1211.5063). arXiv. <http://arxiv.org/abs/1211.5063>
- Peroni, S., & Shotton, D. (2012). FaBio and CiTO: Ontologies for Describing Bibliographic Resources and Citations. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3198992>
- Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, 5(2). <https://doi.org/10.22364/bjmc.2017.5.2.05>
- Pride, D. (2022). *Identifying and Capturing the Semantic Aspects of Citations*. <https://doi.org/10.21954/OU.RO.000146FF>
- Pride, D., & Knoth, P. (2017). Incidental or Influential? - Challenges in Automatically Detecting Citation Importance Using Publication Full Texts. In J. Kamps, G. Tsakonas, Y. Manolopoulos, L. Iliadis, & I. Karydis (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 10450, pp. 572–578). Springer International Publishing. https://doi.org/10.1007/978-3-319-67008-9_48
- Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics*, 118(1), 21–43. <https://doi.org/10.1007/s11192-018-2961-x>
- Ribeiro, M., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for*

Computational Linguistics: Demonstrations, 97–101. <https://doi.org/10.18653/v1/N16-3020>

- Ritchie, A. (2008). *Citation Context Analysis for Information Retrieval*. https://www.researchgate.net/publication/230800457_Citation_Context_Analysis_for_Information_Retrieval
- Robins, R. H. (1970). Ferdinand de Saussure, *Cours de Linguistique Générale*, édition critique par Rudolph Engler, volumes 1–3. Wiesbaden: 1967–1968. Pp. xii + 515. *Journal of Linguistics*, 6(2), 302–304. <https://doi.org/10.1017/S0022226700002711>
- Seewald, A. K. (2002, January). How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness. *Proceedings of the 19th International Conference on Machine Learning*, 554–561.
- Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). *A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2305.17473>
- Small, H. (2018). Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *Journal of Informetrics*, 12(2), 461–480. <https://doi.org/10.1016/j.joi.2018.03.007>
- Smyth, P., & Wolpert, D. H. (1997). Stacked Density Estimation. *Neural Information Processing Systems*.
- Soares, C., Brazdil, P. B., & Kuba, P. (2004). A Meta-Learning Method to Select the Kernel Width in Support Vector Regression. *Machine Learning*, 54(3), 195–209. <https://doi.org/10.1023/B:MACH.0000015879.28004.9b>
- Su, X., Prasad, A., Kan, M.-Y., & Sugiyama, K. (2019). Neural Multi-task Learning for Citation Function and Provenance. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 394–395. <https://doi.org/10.1109/JCDL.2019.00122>
- Sula, C. A., & Miller, M. (2014). Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3), 452–464. <https://doi.org/10.1093/lilc/fqu019>
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, 103. <https://doi.org/10.3115/1610075.1610091>
- Trier, J. O. (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines Sprachlichen Feldes*.
- Trindade Neves, F., Aparicio, M., & De Castro Neto, M. (2024). The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities. *Applied Sciences*, 14(5), 2209. <https://doi.org/10.3390/app14052209>
- Valenzuela-Escárcega, M.-A., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (Version 7). arXiv. <https://doi.org/10.48550/ARXIV.1706.03762>
- Wallin, J. A. (2005). Bibliometric Methods: Pitfalls and Possibilities. *Basic & Clinical Pharmacology & Toxicology*, 97(5), 261–275. https://doi.org/10.1111/j.1742-7843.2005.pto_139.x
- Wu, S., Li, J., & Ding, W. (2023). A geometric framework for multiclass ensemble classifiers. *Machine Learning*, 112(12), 4929–4958. <https://doi.org/10.1007/s10994-023-06406-w>
- Xu, H., Martin, E., & Mahidadia, A. (2013). Using heterogeneous features for scientific citation classification. *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1906.08237>
- Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, 166, 105–126. <https://doi.org/10.1016/j.neunet.2023.06.028>
- Zhao, D., Wang, X., Mu, Y., & Wang, L. (2021). Experimental Study and Comparison of Imbalance Ensemble Classifiers with Dynamic Selection Strategy. *Entropy (Basel, Switzerland)*, 23(7), 822. <https://doi.org/10.3390/e23070822>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). *A Survey of Large Language Models* (Version 13). arXiv. <https://doi.org/10.48550/ARXIV.2303.18223>

- Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., Wen, Y., & Tao, D. (2025). Learning from models beyond fine-tuning. *Nature Machine Intelligence*, 7(1), 6–17. <https://doi.org/10.1038/s42256-024-00961-0>

Appendix

A.1 XAI Experimental Results



Figure A.1. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each specific class of the SciCite dataset in WoS setting. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes).



Figure A.2. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each specific class of the SciCite dataset in WS setting. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes).

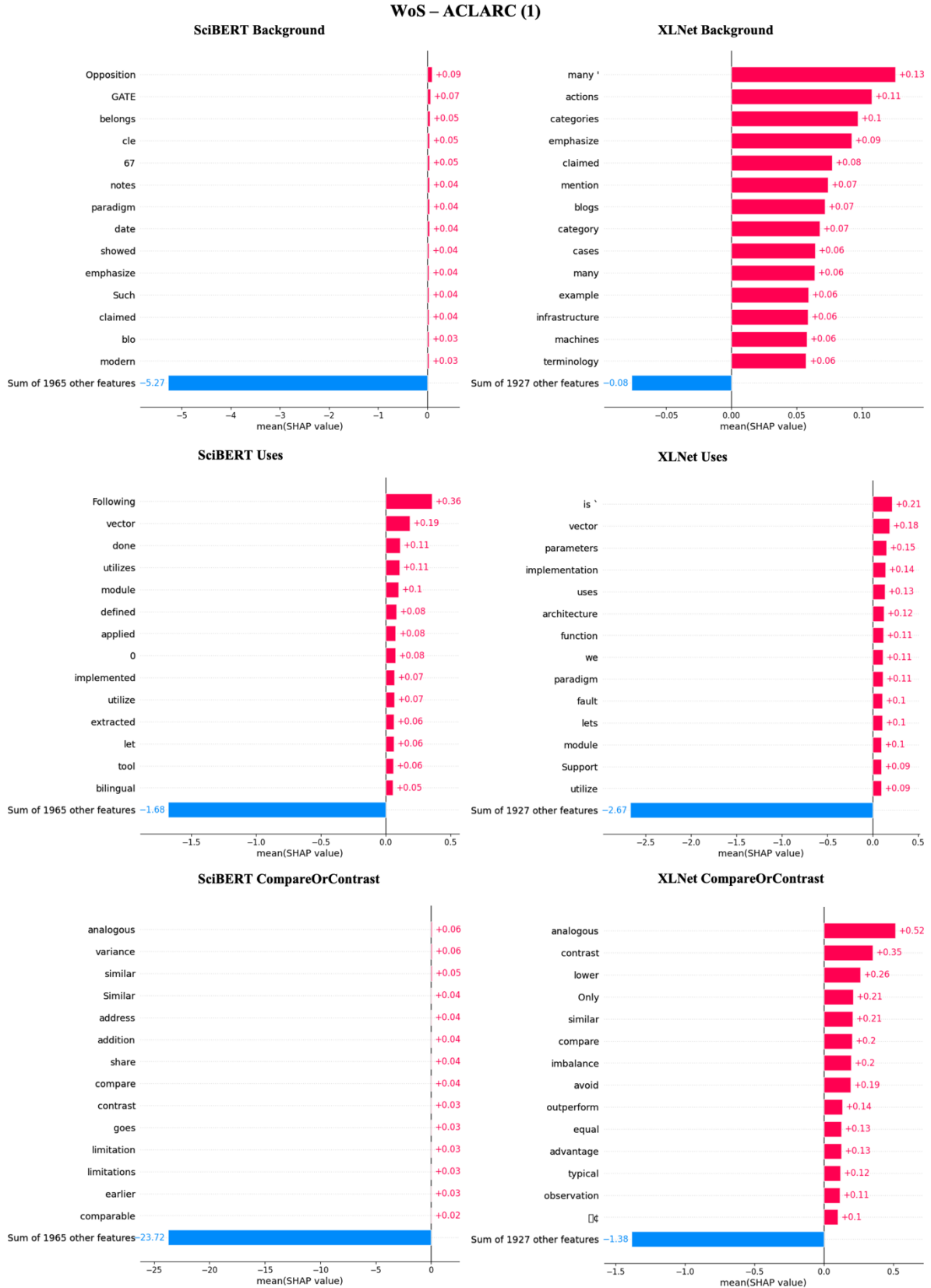


Figure A.3. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each of the first 3 classes (Background, Uses, CompareOrContrast) of the ACL-ARC dataset in WoS setting. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes). Second part in Figure A4.

WoS – ACLARC (2)

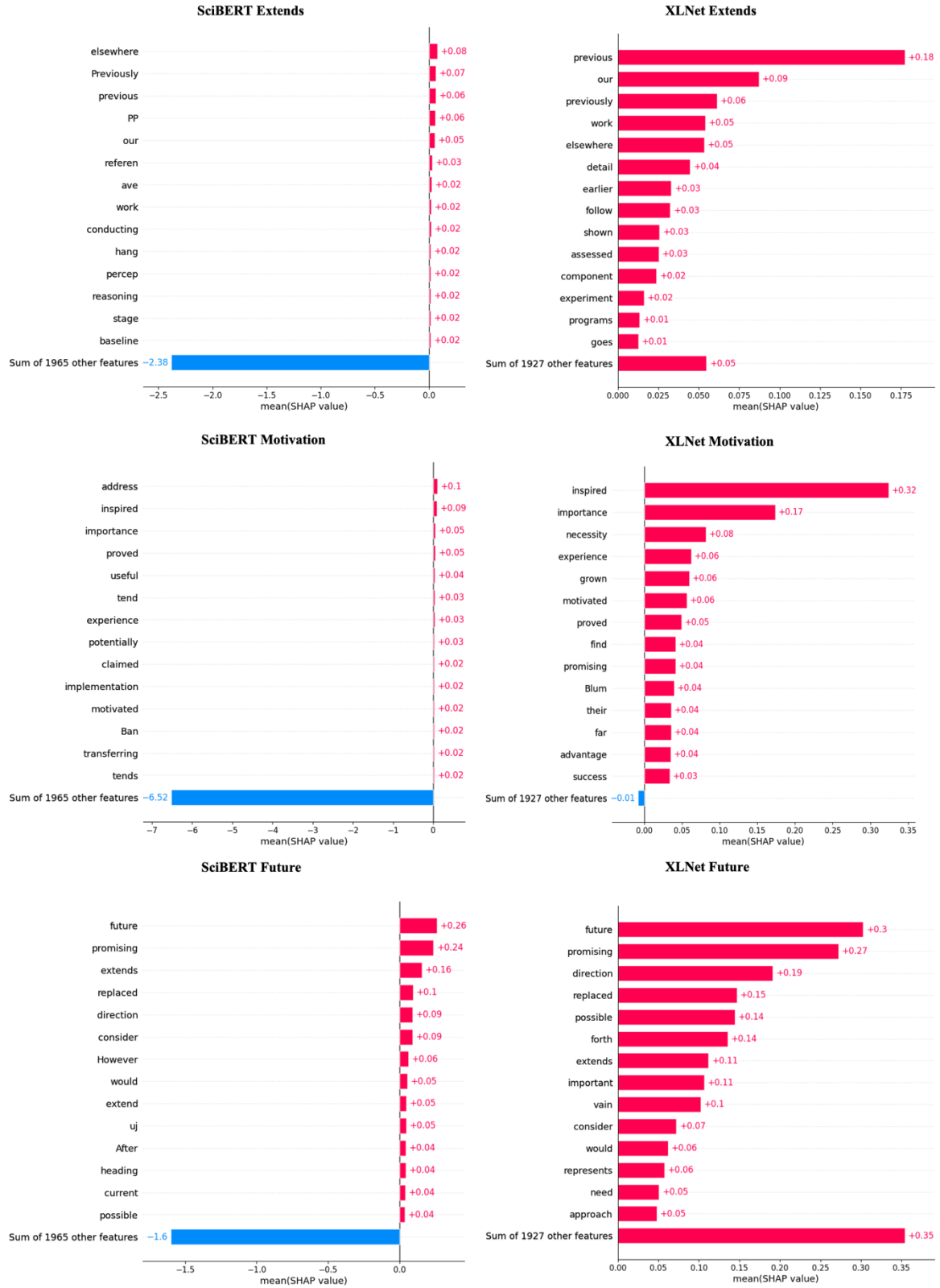


Figure A.4. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each of the remaining 3 classes (Extends, Motivation, Future) of the ACL-ARC dataset in WoS setting, w.r.t. the 3 reported in Figure A3. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes). First part in Figure A3.

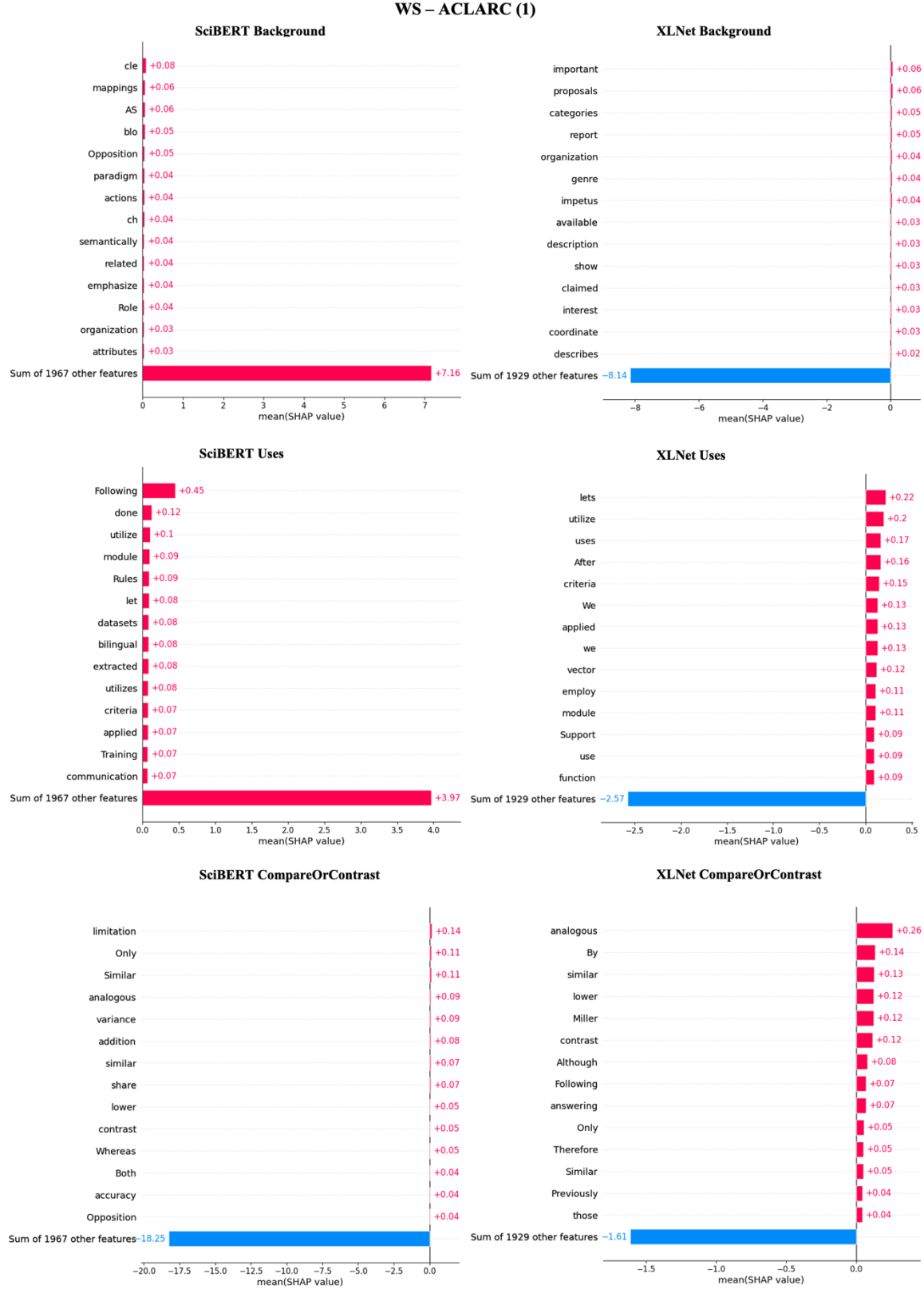


Figure A.5. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each of the first 3 classes (Background, Uses, CompareOrContrast) of the ACL-ARC dataset in WS setting. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes). Second part in Figure A6.

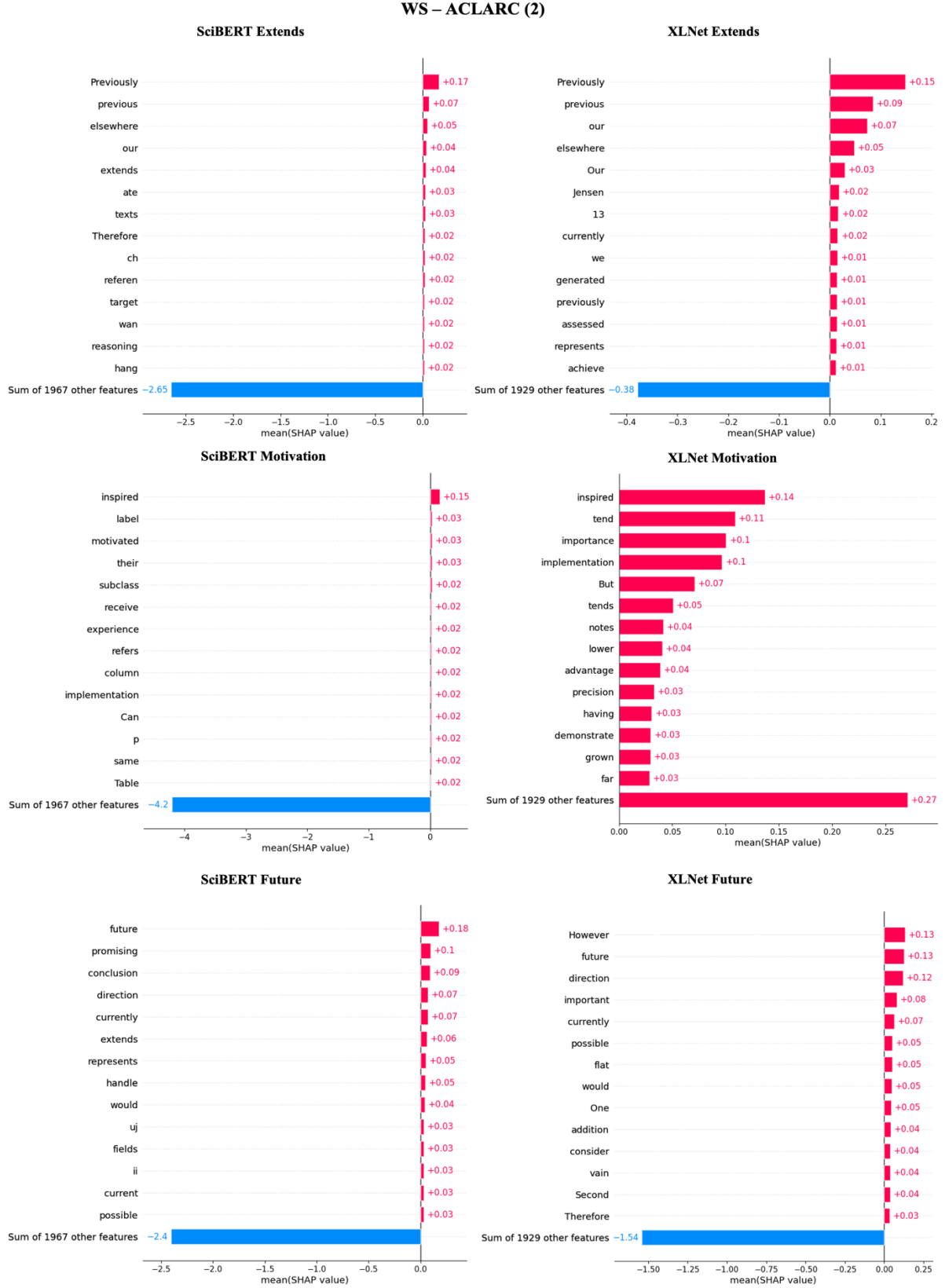


Figure A.6. This figure displays the top 15 tokens, considered as features based on their SHAP scores, that contribute positively to the classification of citations into each of the remaining 3 classes (Extends, Motivation, Future) of the ACL-ARC dataset in WS setting, w.r.t. the 3 reported in Figure A5. Each plot illustrates the contributions of a particular model architecture and the tokens to the classification of the specific class, reported in each title (6 figures: 2 PLMs and 3 classes). First part in Figure A5.

A.2 Computational Instability Details on Minima

SciCite						
Run	SciBERT-based Models			XLNet-based Models		
	Met	Bkg	Res	Met	Bkg	Res
0	0.2385	0.3245	0.1330	0.3139	0.2461	0.1242
1	0.2381	0.3245	0.1331	0.3139	0.2461	0.1242
2	0.2387	0.3245	0.1329	0.3139	0.2461	0.1242
3	0.2386	0.3244	0.1327	0.3139	0.2461	0.1242
4	0.2383	0.3245	0.1328	0.3139	0.2461	0.1242
5	0.2385	0.3245	0.1329	0.3139	0.2461	0.1242
6	0.2385	0.3246	0.1330	0.3139	0.2461	0.1242
7	0.2383	0.3245	0.1330	0.3139	0.2461	0.1242
8	0.2383	0.3245	0.1329	0.3139	0.2461	0.1242
9	0.2382	0.3244	0.1331	0.3139	0.2461	0.1242

Table A.1. Detailed results of computational instability analyses for minima in validation losses within the entire fine-tuning loops recorded for all the level-0 models in each different run (across the 10 total runs) of the same experiment on the SciCite dataset. Models are described by the class on which they were tuned, and by their architecture. Abbreviations reported in the table for these classes are: Met (Method), Bkg (Background), and Res (Result Comparison).

ACL-ARC												
Run	SciBERT-based Models						XLNet-based Models					
	Bkg	Use	CoC	Ext	Mot	Fut	Bkg	Use	CoC	Ext	Mot	Fut
0	0.3173	0.1565	0.2788	0.0872	0.0861	0.0127	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
1	0.3172	0.1574	0.2724	0.0871	0.0916	0.0126	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
2	0.3171	0.1570	0.2738	0.0871	0.0924	0.0126	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
3	0.3171	0.1572	0.2747	0.0873	0.0792	0.0086	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
4	0.3171	0.1573	0.2745	0.0872	0.0792	0.0126	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
5	0.3172	0.1568	0.2745	0.0872	0.0852	0.0132	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
6	0.3172	0.1569	0.2748	0.0872	0.0796	0.0126	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
7	0.3172	0.1566	0.2745	0.0871	0.0788	0.0127	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
8	0.3172	0.1570	0.2747	0.0873	0.0787	0.0084	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075
9	0.3172	0.1574	0.2745	0.0874	0.0792	0.0073	0.3913	0.2195	0.2887	0.1106	0.1745	0.0075

Table A.2. Detailed results of computational instability analyses for minima in validation losses within the entire fine-tuning loops recorded for all the level-0 models in each different run (across the 10 total runs) of the same experiment on the ACL-ARC dataset. Models are described by the class on which they were tuned, and by their architecture. Abbreviations reported in the table for these classes are: Bkg (Background), Use (Uses), CoC (CompareOrContrast), Ext (Extends), Mot (Motivation), and Fut (Future).