Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.2026024160

Unveiling the Mechanisms Between Writing Style and Academic Impact: A Multidimensional and Machine Learning Perspective

Ruixuan Yu*

E-commerce Program, College of Business and Tourism, Sichuan Agricultural University, Sichuan, China

**Corresponding author: Ruixuan Yu*

Abstract

Writing style, as a key non-content factor influencing academic dissemination, remains underexplored in terms of its underlying mechanisms with scientific impact. Drawing on an explainable machine learning perspective, this study investigates the effects of multidimensional writing style features on citation behavior. 23,644 paper abstracts from the Web of Science (WoS) in the fields of computer science, management, and biomedicine were analyzed, extracting four categories of features: readability, emotional and subjective expression, lexical richness, and coherence of expression. These were integrated with 17 baseline indicators to construct an analytical framework. Through ablation experiments, feature importance analysis, and multiple regression, we systematically evaluated the independent contributions and directional effects of each style dimension. The results reveal that readability exerts the strongest predictive power among writing style dimensions and plays the most substantial role in citation behavior. Concise and clear expression significantly enhances citations. Additionally, positive emotional and subjective expressions contribute to improved dissemination outcomes, whereas overly complex vocabulary and excessive cohesion may exert negative effects. This study provides empirical evidence for understanding the relationship between writing style and academic impact, offering practical implications for scholarly writing, journal peer review, and research management.

Keywords

writing style, scientific impact, academic writing, machine learning, science of science

1. Introduction

In academia, the accumulation and dissemination of scientific impact are of paramount importance. They not only determine the efficiency of resource allocation but also profoundly shape the trajectory of knowledge innovation and researchers' career development [1]. Within this process, citation counts have long served as a central indicator of academic impact and remain a focal point of scholarly attention [2, 3]. While the intrinsic quality and originality of research constitute the fundamental prerequisite for high impact [4], mounting evidence indicates that non-content factors—most notably writing style—also play a critical role in academic dissemination [5, 6]. In an era of information overload, effective writing can substantially enhance a paper's discoverability and comprehensibility, thereby directly influencing the depth of its dissemination and the duration of its citation lifespan[7].

Despite the widely acknowledged importance of writing style, existing research still exhibits significant limitations in elucidating its specific mechanisms of influence. Most prior studies have concentrated on isolated dimensions—such as readability [8] or emotional and subjective expression [9]—and have generally reported positive correlations between these features and citation rates. However, such approaches often reduce the complex, multidimensional construct of writing style to independent variables, overlooking its nature as an integrated, synergistic system. Under this fragmented analytical framework, the field lacks a systematic understanding of how writing style collectively shapes academic impact, which in turn leaves scholars without robust empirical guidance for writing practice.

A critical question therefore emerges: Among the various dimensions of writing style, which exert the greatest influence on academic dissemination? Although previous work has highlighted the roles of readability, emotional valence, and other features, findings remain scattered and at times contradictory. For instance, some studies advocate for concise and objective narratives as more highly valued [10], while others suggest that moderate subjectivity can enhance persuasiveness [11]. Such inconsistencies create practical confusion for scholars: given limited time and effort, which aspects should be prioritized for optimization? Systematically clarifying the relative contributions of different writing style dimensions to citation impact—and thereby identifying the most decisive drivers—thus represents an urgent and practically significant task for refining dissemination strategies and elevating academic influence.

Furthermore, recognizing the importance of a feature is only half the story; we must also decipher how it exerts its influence. Even if we can identify which stylistic elements matter most, without understanding the precise pathways through which they operate, actionable writing recommendations remain elusive. For example, does greater lexical diversity lead to more citations [12], or does excessive complexity create barriers [13]? Is clear and accessible prose more conducive to knowledge dissemination[14], or does deliberately obscure writing confer greater perceived academic authority [15, 16]? The “black-box” nature of many predictive models often conceals the direction and nature of these relationships [17]. Clarifying the exact associations between writing style dimensions and citation behavior is therefore essential for developing evidence-based guidelines for scholarly writing.

To address these challenges, this study constructs an interdisciplinary empirical framework to systematically examine the mechanisms through which multidimensional writing style structures influence academic impact. We selected authoritative journals in computer science, management, and biomedicine, compiling a corpus of 23,644 paper abstracts. Through multidimensional quantitative analysis, four core features were extracted: readability, emotional and subjective expression, lexical richness, and coherence of expression. Machine learning models were then employed to predict citation counts based on these features. Ablation experiments confirmed the significant contributions of all four dimensions to predictive performance. Feature importance analysis revealed readability as the dominant predictor, followed by lexical richness, with emotional expression and coherence also exerting notable effects. Finally, interpretable regression techniques were applied to uncover the specific directional mechanisms of each feature. The results yield nuanced insights: although readability is paramount, its coefficient is negative, indicating that more readable (structurally simpler) abstracts tend to receive higher citations. Similarly, the negative coefficient for lexical richness suggests that simplicity outperforms ornate vocabulary.

The theoretical contributions of this study are threefold. First, by adopting a multidimensional, interactive perspective, we systematically investigate the synergistic effects of writing style elements on citation behavior, overcoming the limitations of prior studies that examined isolated linguistic features. Second, through the application of explainable machine learning techniques, we reveal the internal pathways by which different writing styles influence citation rates, providing new causal evidence for the relationship between writing style and academic impact. Third, by identifying differentiated impact patterns across writing dimensions, we deepen the understanding of the complex interplay between academic writing and knowledge dissemination, laying a theoretical foundation for future research.

The remainder of the paper is organized as follows: Section 2 systematically reviews the relevant literature on academic impact evaluation, factors influencing citation behavior, academic writing quality and stylistic features, and the application of machine learning in scientometrics, providing the theoretical foundation for this study. Section 3 describes the data and methodology. Section 4 presents the empirical results. Section 5 summarizes the findings, discusses their theoretical and practical implications, and outlines limitations and future directions. Section 6 concludes the paper.

2. Related Work

2.1 Scholarly Impact and Influencing Factors

The evaluation of scholarly impact has long been a central topic in academia, serving as a basis for measuring the contributions of researchers, papers, teams, and institutions, and informing decisions on funding allocation, hiring, and awards [1]. Common indicators of scholarly impact include Field-Weighted Citation Impact (FWCI) [18], download counts [19], and others, among which citation counts remain the most widely used and accepted metric [3, 20, 21].

Citation counts are shaped by a multitude of multidimensional factors. Prior research has shown that the number of authors, title length, article type (e.g., research articles versus reviews), and international collaboration all exert significant effects on citations [22]. In addition, article length and number of references [23], keyword count and diversity [24, 25], publication in special issues [26], and open access status [23] are positively correlated with citation rates. Earlier publication years, by contrast, tend to be associated with lower citation counts [27]. Research funding levels [28], content novelty [29], and the use of institutional email addresses [30] have also been confirmed to correlate closely with citation volume. At the same time, systematic differences exist across disciplines, with journals in different fields exhibiting distinct patterns in publication volume and citation accumulation [31, 32]. Usage frequency is likewise crucial: Chi and Glänzel [33] found a significant correlation between citation counts and usage counts in Web of Science. Markusova et al. [34] collected "usage counts in the last 180 days" and "usage counts since 2013" and demonstrated a significant positive relationship between usage indicators and citation indicators. In recent years, academic writing style has increasingly been recognized as a potentially key variable influencing scholarly impact, warranting deeper investigation.

2.2 Writing Style of Academic Papers

Academic writing style constitutes a major topic in scientometrics and scholarly communication research. Its influence can be examined across multiple dimensions, including readability, emotional and subjective expression, lexical richness, and coherence of expression[5, 6, 35].

Regarding readability, McCannon [5] investigated the relationship between readability and citation counts, finding that extremely difficult-to-read articles significantly reduce citations. Lei and Yan [36] verified the role of abstract readability in scientific impact using composite indicators. Dowling et al. [8] analyzed papers published in *Economics Letters* and further demonstrated a positive correlation between readability scores and citation rates. In terms of emotional and subjective expression, Sienkiewicz and Altmann [6] noted that positive emotional polarity increases the likelihood of a paper becoming highly cited. Thelwall et al. [9] observed that, across multiple disciplines, the explicit use of first-person pronouns to state the researcher's position and contributions has become increasingly common and is often associated with higher-quality research. With respect to lexical richness, Lu et al. [37] constructed a writing quality evaluation system based on lexical complexity and syntactic complexity and explored its relationship with scientific impact. Sienkiewicz and Altmann [6] likewise found that lexical diversity (z-index) and Gunning Fog index in abstracts are positively correlated with citation performance—indicating that more diverse vocabulary and complex sentence structures are associated with higher citations. Concerning coherence of expression, Alyousef [35] emphasized the critical role of cohesive devices in enhancing academic text quality. Souza [38] proposed a four-dimensional framework for analyzing discourse coherence in academic abstracts, offering theoretical support for optimizing scientific writing.

Writing style plays a pivotal role in knowledge dissemination and impact construction. However, existing studies have largely focused on single dimensions, treating writing features as independent variables and overlooking the complex interactions among them. To address this gap, the present study introduces machine learning methods to integrate the four linguistic features outlined above and systematically examine the synergistic mechanisms through which academic writing style influences paper impact.

2.3 Machine Learning and Interpretable Machine Learning

Machine learning models demonstrate significant advantages in handling complex multidimensional relationships and have been successfully applied in scientometrics. For example, Rabby and Berka [39] used random forest models to perform multi-classification tasks on COVID-19 biomedical literature and achieved optimal performance. Liu et al. [40] employed six machine learning models for feature extraction and successfully predicted paper retraction probability using XGBoost. Alohali et al. [41] systematically explored the relationship between textual features of papers in otorhinolaryngology and citation frequency across linear regression, boosted decision trees, decision trees, and neural networks. Gradient boosting models and AdaBoost have also proven effective in scientific prediction tasks [42, 43]. In the domain of academic impact prediction, Yan et al. [44] applied SVR and linear regression to forecast future paper influence. Wang et al. [45] constructed a machine learning framework based on content topic features and validated the effectiveness of SVM, KNN, and Bagging classifiers.

Although these traditional machine learning models perform well on complex tasks, their “black-box” nature limits the interpretability of results to some extent. With the development of explainable machine learning—such as SHAP values [46], feature importance ranking [47], and logistic regression [48]—effective tools have emerged for revealing the internal decision-making mechanisms of models and enhancing understanding and trust in predictions. Building on this, the present study employs multiple machine learning models for predictive analysis and utilizes logistic regression together with feature importance ranking to systematically investigate the influence mechanisms of each feature on citation counts, thereby clearly elucidating the intrinsic relationships between writing style features and citation behavior.

3. Data and Methodology

3.1 Data Collection and Preprocessing

The data for this study were sourced from the Web of Science Core Collection database (webofscience.clarivate.cn), with the search conducted on September 7, 2025, covering publications from 2015 to 2025. To enhance the generalizability of the findings, the author collected paper abstracts from disciplines with diverse backgrounds. Specifically, four authoritative journals were selected from each of three fields: computer science (Nature Machine Intelligence, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Computers, Journal of the ACM); management (Management Science, Production and Operations Management, MIS Quarterly, Journal of Operations Management); and biomedicine (Cell Reports Medicine, Nature Biomedical Engineering, IEEE Transactions on Biomedical Engineering, Briefings in Bioinformatics). The initial retrieval yielded 23,710 papers. For each paper, complete metadata and textual content were extracted, including document type, authors, title, abstract, number of references, citation count, funding information, corresponding address, affiliations, author email, publication year, and open access status.

Subsequent preprocessing was performed as follows. To ensure comparability of citation behavior, only three document types were retained: Articles, Reviews, and Editorial Materials. Errata, retractions, and other types with systematically different citation motives from original research were excluded. Data mining was then applied to the included records. Baseline features and writing style features were extracted based on the exported WoS data (see Section 3.2 for details). Finally, duplicate records were removed based on DOI, and entries with completely missing key information (abstract, authors, publication year, etc.) were deleted. After systematic data cleaning, a final valid sample of 23,644 records was obtained for subsequent analysis.

3.2 Variable Measurement

3.2.1 Baseline Measurement

To establish a robust baseline that accounts for well-documented non-linguistic factors influencing citation counts, a set of control variables was derived from each publication’s metadata. The selection of these baseline features was guided by prior scientometric literature. Core fields were extracted from the raw metadata, and additional baseline features were constructed and expanded through data mining and feature engineering—for example, identifying the presence of funding agency text in the “Funding Orgs” field to create a funding indicator, and detecting special issue publication from the "Special Issue" field. Ultimately, 17 baseline features were extracted and processed, with detailed definitions as follows:

(1) citedReferenceCount: Total number of references in the paper (Antoniou et al., 2015), exported directly from the WoS Core Collection.

(2) usageCount_180: Number of accesses or saves in the past 180 days (Markusova et al., 2018), exported from WoS.

(3) usageCountSince2013: Cumulative accesses or saves since 2013 (Markusova et al., 2018), exported from WoS.

(4) pubYear: Publication year (Padial et al., 2010), exported from WoS.

(5) pages: Total number of pages (Antoniou et al., 2015), exported from WoS.

(6) authorNum: Total number of authors, calculated by splitting the author field on semicolons [22].

(7) funding: Funding indicator; coded as 1 if funding information exists in the "Funding Orgs" field, otherwise 0 [28].

(8) openAccess: Open access status; coded as 1 if the "Open Access Designations" field is non-empty, otherwise 0 [23].

(9) collaboration: International collaboration indicator; countries in author addresses were standardized using the pycountry library (e.g., "UK," "England," and "United Kingdom" mapped to "UK"; "America" to "USA"; etc.). Coded as 1 if authors are from multiple countries, otherwise 0 [22].

(10) novelty: Novelty score, computed as the inverse of the average cosine similarity between the paper's TF-IDF vectorized abstract and those of all other papers [29].

(11) titleWords: Number of words in the title, obtained by splitting the title text on spaces [22].

(12) sp: Special issue indicator; coded as 1 if the paper belongs to a special issue, otherwise 0 [26].

(13) keywords: Number of keywords, calculated by splitting the "Keywords Plus" field (missing values imputed as 0) [24, 25].

(14) pubVolume: Annual publication volume of the journal in the paper's publication year, matched from pre-collected annual totals for the 12 journals in the sample (2015-2025) based on Source Title and Publication Year [32].

(15) docType: Document type, mapped to categorical codes: 1 for research articles, 2 for reviews, 3 for editorial materials [22].

(16) subjects: Disciplinary classification; journals were assigned to computer science (1), management (2), or biomedical engineering (3) based on Source Title [31].

(17) emailType: Email type; coded as 0 if the corresponding author's email uses an educational/institutional domain, otherwise 1 [30].

3.2.2 Writing Style Measurement

To isolate the independent effects of writing style on citations while controlling for non-linguistic factors, four categories of linguistic style features were extracted from the abstracts. Feature selection and measurement were grounded in existing literature on academic writing and scientific communication. The four dimensions, along with their definitions and computation methods, are as follows:

(1) Readability

Readability was quantified using the SMOG index (Simple Measure of Gobbledygook), which assesses text complexity based on the number of polysyllabic words and sentence length. It is widely used in scholarly studies [49, 50] Higher SMOG scores indicate lower readability (more complex text). The SMOG formula is:

$$SMOG=1.043\times\sqrt{\left(\left(poly\text{-}syllables\times 30\right)/\text{sentences}\right)}+3.1291 \quad (1)$$

where *poly-syllables* is the count of words with three or more syllables (computed using pyphen on content words), and *sentences* is the total number of sentences (obtained via the textstat library in Python).

(2) Emotional and Subjective Expression (emoExpression)

Emotional polarity (positive-negative valence) and subjectivity (degree of opinionated/personal expression) were used to quantify affective and subjective tone. These metrics analyze lexical sentiment and the strength of subjective viewpoints and are commonly applied in stylistic text analysis [51, 52]. Both were computed using Python's TextBlob library. Higher values indicate more positive emotion and greater subjectivity.

(3) Lexical Richness

Lexical richness was assessed using three complementary indicators: MTLD (Measure of Textual Lexical Diversity), Lexical Density, and Syntactic Density. MTLD calculates the average length of word sequences before repetition and is regarded as a robust measure of lexical diversity [53]. Lexical Density is the proportion of content words (nouns, verbs, adjectives, adverbs) to total words, reflecting informational richness [54]. Syntactic Density is the ratio of clauses to sentences, capturing syntactic complexity [55]. Part-of-speech tagging was performed with spaCy, and metrics were computed using the lexical-richness library in Python. Higher values indicate greater lexical diversity, informational density, and syntactic complexity, respectively.

(4) Coherence of Expression

Coherence was measured through semantic overlap between adjacent sentences (Sem1_avg), which captures semantic continuity in discourse [56]. Higher Sem1_avg values indicate greater semantic coherence. The computation is as follows:

Coherence was measured through semantic overlap between adjacent sentences (Sem1_avg), which captures semantic continuity in discourse [56]. Higher Sem1_avg values indicate greater semantic coherence. The computation is as follows:

$$Sem1_{a}vg=(1/(n-1))\times\Sigma cosine_{s}imilarity(v_{i},v_{i-1}) \qquad (2)$$

where v_i is the semantic vector of the i-th sentence (obtained via TF-IDF vectorization followed by TruncatedSVD dimensionality reduction to 100 dimensions using scikit-learn); v_{i-1} is the semantic vector of the previous sentence; $cosine\_similarity(v_i,v_{i-1})$ is the cosine similarity between adjacent sentence vectors; and n is the total number of sentences in the abstract. The average similarity across all consecutive sentence pairs measures overall semantic coherence.

3.3 Research Framework

To systematically investigate the predictive power of different feature combinations on citation counts (timesCited), this study designed multiple feature ablation experiments based on the following six combinations:

(1) Baseline features only: the 17 non-linguistic control variables listed in Section 3.2.1.

(2) Baseline + readability: adding the SMOG index.

(3) Baseline + emotional and subjective expression: adding polarity and subjectivity.

(4) Baseline + lexical richness: adding MTLD, lexical density, and syntactic density.

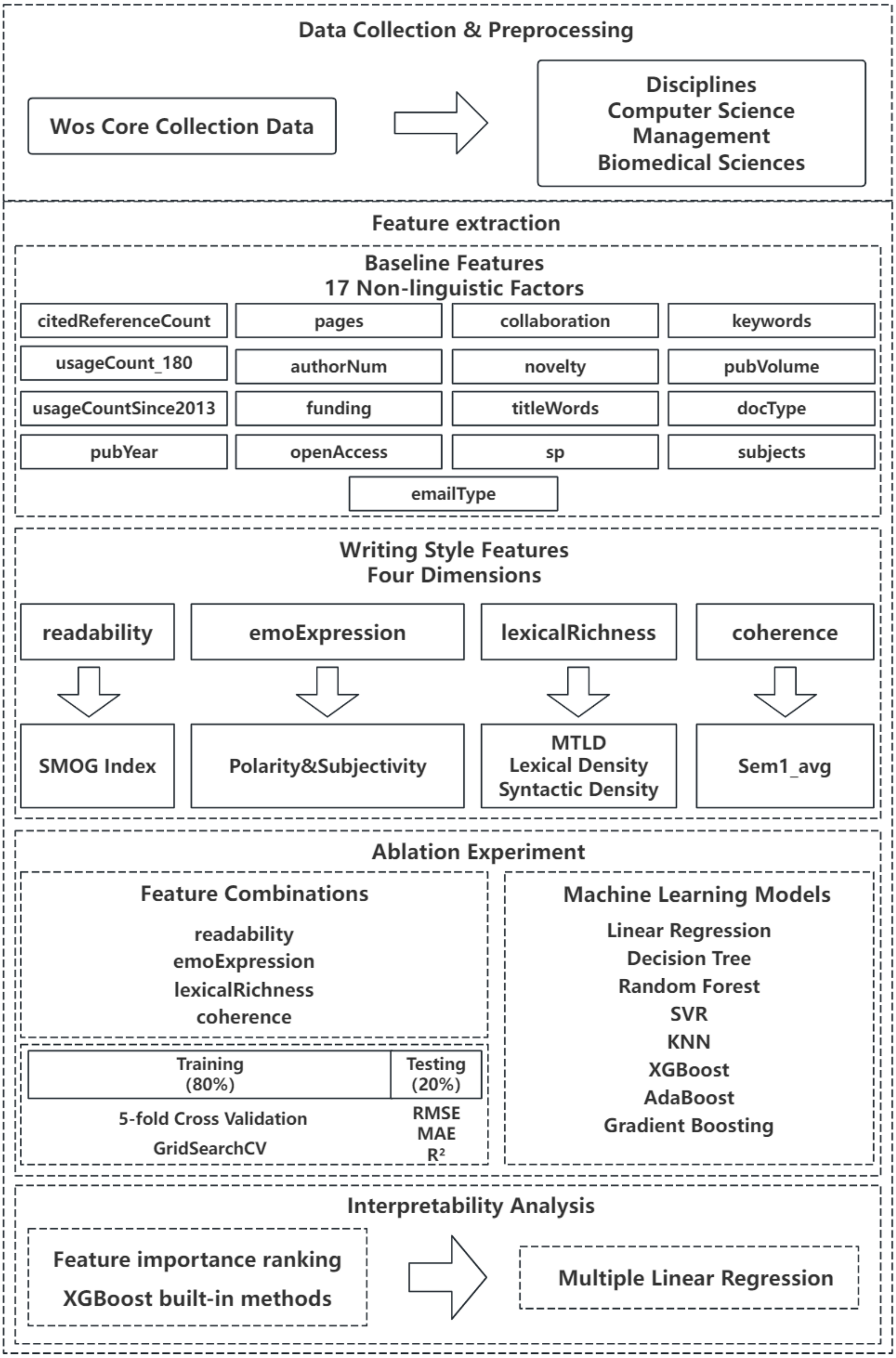
(5) Baseline + coherence: adding the semantic coherence indicator (Sem1_avg).

(6) Full feature set: combining all four writing style categories with the baseline features.

For model selection, eight mainstream regression algorithms were compared: Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), XGBoost, AdaBoost, and Gradient Boosting. All models were trained and tested with citation count as the target variable.

To identify the optimal predictive model, a rigorous evaluation pipeline was followed. The dataset was randomly split into training (80%) and test (20%) sets. GridSearchCV with five-fold cross-validation was applied to optimize hyperparameters for each model, using Root Mean Square Error (RMSE) as the evaluation criterion to determine the best configuration within predefined parameter grids [57]. Final model performance was comprehensively assessed using RMSE, Mean Absolute Error (MAE), and the coefficient of determination (R²) to identify the best-performing model under each feature combination. The overall research process is illustrated in Figure 1.

Figure 1: Research Framework of This Study



4. Research Result

This study systematically examined the predictive contribution, relative importance, and specific directional effects of writing style features on paper citation counts through empirical analysis. First, ablation experiments were conducted by incrementally adding the four categories of writing style features (readability, emotional and subjective expression, lexical richness, and coherence of expression) to the 17 baseline features, with predictive performance evaluated across eight machine learning regression models. Second, to further clarify the relative importance of each feature, feature importance ranking was performed using the optimal predictive model after standardizing and aggregating all features. Finally, to elucidate the precise relationships between writing style dimensions and citation behavior, multiple linear regression models were constructed to reveal the direction and statistical significance of each feature's effect. All models underwent hyperparameter optimization via grid search, with performance evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

4.1 Ablation Experiment Results: Baseline

When using only baseline features, the Random Forest model achieved the best predictive performance (RMSE = 119.57, R^2 = 0.325), as shown in Table 1. This result reflects, to some extent, the foundational predictive capacity of non-linguistic factors such as number of authors, references, and funding in explaining citation counts. However, these non-linguistic factors alone exhibit certain limitations and fail to fully capture the potential influence of writing style on academic dissemination. Accordingly, this baseline model (Random Forest performance on baseline features) serves as the reference benchmark. Subsequent analyses introduce writing style features to assess their marginal contributions relative to the baseline, thereby providing deeper insight into the role of writing style in citation behavior.

Table 1: Predictive Results Using Baseline Features

Model	Best_Params	RMSE	MAE	R2
Random Forest	{‘model__max_depth’:30, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:5, ‘model__n_estimators’: 100}	**119.57**	**23.50**	**0.325**
Decision Tree	{‘model__max_depth’:10, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 5}	121.26	25.61	0.306
GradientBoosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:6, ‘model__n_estimators’:100, ‘model__subsample’: 1.0}	128.30	23.81	0.223
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:7, ‘model__weights’: ‘uniform’}	129.30	28.41	0.211
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.37	27.85	0.210
AdaBoost	{‘model__learning_rate’:0.01, ‘model__n_estimators’: 50}	134.44	40.92	0.147
XGBoost	{‘model__learning_rate’:0.1, ‘model__max_depth’:3, ‘model__n_estimators’:200, ‘model__subsample’: 1.0}	159.97	27.29	-0.208
Linear Regression	No hyperparameter tuning	205.42	87.17	-0.992

4.2 Ablation Experiment Results: Readability

The inclusion of readability features significantly improved overall model performance. As shown in Table 2, the Gradient Boosting model performed best under this feature combination (RMSE = 110.10, R^2 = 0.428), with RMSE reduced by 9.47 units (a 7.9% improvement) compared to the best baseline model. Across all models, R^2 values were positive and generally higher than those of their baseline counterparts, indicating that text readability provides stable and substantial predictive value for paper impact.

Table 2: Predictive Results with Readability Features

Model	Best_Params	RMSE	MAE	R2
GradientBoosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:6, ‘model__n_estimators’:200, ‘model__subsample’: 0.8}	**110.10**	**22.67**	**0.428**
XGBoost	{‘model__learning_rate’:0.1, ‘model__max_depth’:3, ‘model__n_estimators’:200, ‘model__subsample’: 0.8}	112.70	24.99	0.400
Random Forest	{‘model__max_depth’:20, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:2, ‘model__n_estimators’: 100}	113.69	23.40	0.390
Decision Tree	{‘model__max_depth’:20, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 2}	116.53	29.66	0.359
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:3, ‘model__weights’: ‘uniform’}	125.45	30.69	0.257
AdaBoost	{‘model__learning_rate’:0.01, ‘model__n_estimators’: 50}	128.60	40.61	0.219
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.36	27.84	0.210
Linear Regression	No hyperparameter tuning	201.53	87.23	-0.917

4.3 Ablation Experiment Results: emoExpression

Emotional and subjective expression features also contributed predictive information to citation counts (Table 3). The Gradient Boosting model again achieved the best performance in this combination (RMSE = 117.96, R^2 = 0.343). Notably, while the performance gain from adding emotional and subjective expression features was smaller than that from readability, this may be attributable to the inherently restrained nature of emotional expression in academic texts [58]. The limited variability in emotional expression results in a relatively modest overall contribution to prediction.

Table 3: Predictive Results with Emotional and Subjective Expression Features

Model	Best_Params	RMSE	MAE	R2
GradientBoosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:6, ‘model__n_estimators’:200, ‘model__subsample’: 0.9}	**117.96**	**22.79**	**0.343**
Random Forest	{‘model__max_depth’:30, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:2, ‘model__n_estimators’: 100}	118.89	23.72	0.333
Decision Tree	{‘model__max_depth’:10, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 5}	121.02	25.40	0.309
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.34	27.83	0.210
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:7, ‘model__weights’: ‘uniform’}	130.40	29.69	0.197
AdaBoost	{‘model__learning_rate’:0.01, ‘model__n_estimators’: 50}	135.95	41.19	0.127
XGBoost	{‘model__learning_rate’:0.05, ‘model__max_depth’:3, ‘model__n_estimators’:200, ‘model__subsample’: 1.0}	161.62	28.21	-0.233
Linear Regression	No hyperparameter tuning	205.76	87.29	-0.999

4.4 Ablation Experiment Results: Lexical Richness

Lexical richness features likewise demonstrated solid predictive capability (Table 4). Under this feature combination, the XGBoost model delivered the best performance (RMSE = 117.15, R^2 = 0.352). Compared to the baseline benchmark, performance improved, indicating that indicators such as lexical diversity, lexical density, and syntactic complexity offer explanatory power for citation behavior. This result likely stems from the ability of lexical richness metrics to quantify the lexical and syntactic complexity of abstracts, which are closely tied to information density and expression style [59] and thus statistically predict citation counts. Although the predictive gain from this dimension was lower than that from readability, it exceeded the contribution of emotional and subjective expression, suggesting a meaningful independent role within the multidimensional analysis of writing style.

Table 4: Predictive Results with Lexical Richness Features

Model	Best_Params	RMSE	MAE	R2
XGBoost	{‘model__learning_rate’:0.1, ‘model__max_depth’:6, ‘model__n_estimators’:300, ‘model__subsample’: 1.0}	117.15	23.45	0.352
Random Forest	{‘model__max_depth’:None, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:2, ‘model__n_estimators’: 100}	123.00	23.71	0.286
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.37	27.85	0.210
GradientBoosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:9, ‘model__n_estimators’:300, ‘model__subsample’: 0.8}	129.86	23.48	0.204
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:9, ‘model__weights’: ‘distance’}	134.55	29.81	0.145
AdaBoost	{‘model__learning_rate’:0.01, ‘model__n_estimators’: 100}	139.76	43.68	0.078
Decision Tree	{‘model__max_depth’:20, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 10}	151.18	28.50	-0.079
Linear Regression	No hyperparameter tuning	203.87	88.31	-0.962

4.5 Ablation Experiment Results: Coherence

Coherence of expression features also exhibited stable predictive value (Table 5). The Gradient Boosting model achieved the best performance in this combination (RMSE = 116.77, $R^2 = 0.356$). Model comparisons showed that the addition of this feature improved predictive accuracy more than lexical richness or emotional expression features, ranking second only to readability among writing style dimensions. This result may be attributed to the ability of coherence features to effectively capture semantic connectivity between sentences in abstracts [60]. By measuring semantic similarity across adjacent sentences, the indicator reflects logical fluency—a key factor influencing reader comprehension efficiency—and thus demonstrates strong explanatory power in citation prediction.

Table 5: Predictive Results with Coherence Features

Model	Best_Params	RMSE	MAE	R2
Gradient Boosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:6, ‘model__n_estimators’:300, ‘model__subsample’: 0.8}	116.77	22.72	0.356
Random Forest	{‘model__max_depth’:10, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:5, ‘model__n_estimators’: 100}	121.22	23.84	0.306
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.44	27.85	0.209
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:7, ‘model__weights’: ‘uniform’}	129.60	29.13	0.207
AdaBoost	{‘model__learning_rate’:0.01, ‘model__n_estimators’: 50}	135.86	41.29	0.129
XGBoost	{‘model__learning_rate’:0.05, ‘model__max_depth’:3, ‘model__n_estimators’:200, ‘model__subsample’: 1.0}	160.25	28.29	-0.212
Decision Tree	{‘model__max_depth’:None, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 2}	203.43	33.16	-0.954
Linear Regression	No hyperparameter tuning	207.82	87.80	-1.039

4.6 Ablation Experiment Results: All Four Language Features + Baseline

When all baseline features were combined with the four categories of writing style features, overall predictive performance improved further. As shown in Table 6, the XGBoost model performed best on this comprehensive feature set (RMSE = 111.33, $R^2 = 0.415$). Interestingly, while the full model outperformed most single-dimension feature combinations, its predictive accuracy was slightly lower than that of the model using only readability features (RMSE = 110.10). This phenomenon suggests that readability possesses stronger predictive dominance among writing style dimensions and tends to overshadow other features in the overall predictive process [61]. Nevertheless, the other writing style dimensions still provided marginal contributions, and their context-specific value warrants further exploration through more granular analytical approaches.

Table 6: Ablation Experiment Results for Writing Style Features

Model	Best_Params	RMSE	MAE	R2
XGBoost	{‘model__learning_rate’:0.1, ‘model__max_depth’:3, ‘model__n_estimators’:300, ‘model__subsample’: 1.0}	111.33	24.68	0.415
Random Forest	{‘model__max_depth’:None, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’:2, ‘model__n_estimators’: 200}	118.48	23.75	0.337
GradientBoosting	{‘model__learning_rate’:0.05, ‘model__max_depth’:6, ‘model__n_estimators’:200, ‘model__subsample’: 0.9}	122.51	23.08	0.291
KNN	{‘model__algorithm’:’auto’, ‘model__n_neighbors’:3, ‘model__weights’: ‘uniform’}	127.32	32.74	0.235
SVR	{‘model__C’: 100, ‘model__gamma’: ‘scale’, ‘model__kernel’: ‘linear’}	129.37	27.83	0.210
Decision Tree	{‘model__max_depth’:10, ‘model__min_samples_leaf’:1, ‘model__min_samples_split’: 5}	129.88	26.73	0.204
AdaBoost	{‘model__learning_rate0.01, ‘model__n_estimators’: 50}	132.34	42.15	0.173
Linear Regression	No hyperparameter tuning	203.49	88.46	-0.955

4.7 Summary of Ablation Experiments

To systematically evaluate the incremental predictive contributions of the four writing style categories to citation counts, ablation experiments were conducted by sequentially adding each language feature category to the baseline features. The results demonstrate that the inclusion of any writing style module improves model predictive accuracy, confirming the effectiveness of linguistic features in academic impact prediction. As shown in Table 7, compared to the baseline model (RMSE = 119.57, $R^2 = 0.325$), the best-performing model under each writing style dimension exhibited reduced RMSE. Among them, readability contributed the most substantially, lowering RMSE by 9.47 units (a 7.9% improvement). Emotional and subjective expression, lexical richness, and coherence features each brought varying degrees of performance enhancement. The ablation experiments collectively affirm that academic writing style is an important factor influencing citation behavior, and its incorporation into predictive models effectively enhances performance.

Table 7: Best-Performing Models for Each Feature in Ablation Experiments

Writing style features	Best Model	RMSE	MAE	R2
Baseline	Random Forest	119.57	23.50	0.325
readability	GradientBoosting	117.96	22.79	0.343
emoExpression	GradientBoosting	117.96	22.79	0.343
lexicalRichness	XGBoost	117.15	23.45	0.352
coherence	GradientBoosting	116.77	22.72	0.356

4.8 Interpretable Machine Learning

4.8.1 Standardization and Aggregation of Writing Style Features

To accurately assess the relative importance of different writing style dimensions on paper impact, a composite linguistic feature set was constructed. This set was formed by standardizing the indicators from the four writing style dimensions and integrating them. Given differences in the original scales of the linguistic features, which could introduce assessment bias, all linguistic features were standardized using Scikit-learn’s StandardScaler [62] to achieve a mean of 0 and standard deviation of 1, ensuring comparability across dimensions. Predictive performance was then re-evaluated across the eight machine learning models on this standardized feature set. As shown in Table 8, the XGBoost model performed best (RMSE = 112.836, $R^2 = 0.399$) and was therefore selected as the base model for subsequent feature importance analysis.

Table 8: Best Model Evaluation Results

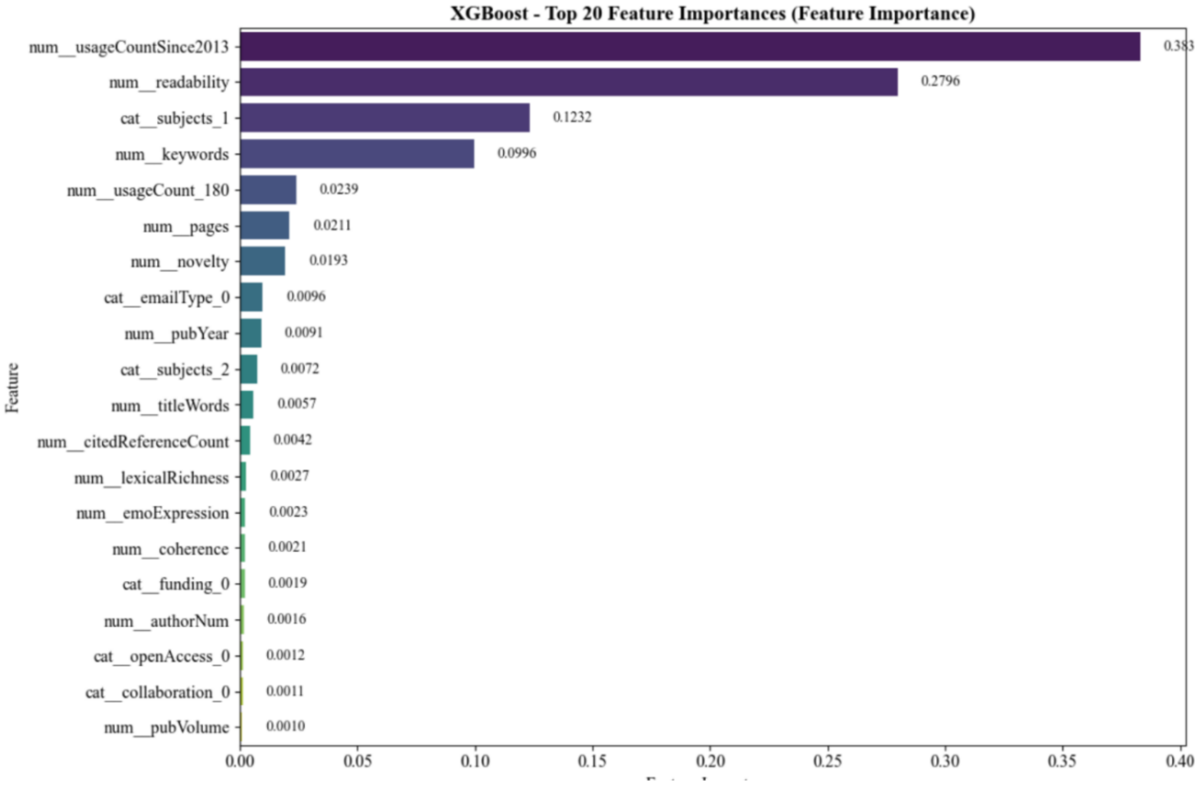
Model Name	Best Parameters	RMSE	MSE	MAE	R^2
XGBoost	{'model__learning_rate':0.05, 'model__max_depth':3, 'model__n_estimators':300, 'model__subsample': 1.0}	112.8360	12731.9737	25.3241	0.3989

4.8.2 Feature Importance Ranking

After selecting the best-performing XGBoost model as the base model, feature importance ranking was conducted to evaluate the relative contributions of each writing style dimension to citation counts. Feature importance scores reflect a feature's relative usefulness in the prediction task and serve as a primary step in opening the "black box" of machine learning models. The method computes importance by aggregating the frequency with which a feature is used as a splitting node across all decision trees [63]; it is widely adopted due to its scalability and stability in handling large-scale data and complex models [64]. In this study, feature importance was obtained directly from the trained XGBoost model's feature_importances_ attribute.

The feature importance ranking results are presented in Figure 2. Historical usage count (usageCountSince2013) emerged as the most important predictor (importance = 0.383), followed by readability (importance = 0.280). Disciplinary field and keyword count ranked third and fourth, respectively. The top four features accounted for a cumulative importance of 0.886, exhibiting a classic "vital few" distribution. Among the four writing style categories, readability demonstrated the strongest predictive power, while lexical richness, emotional expression, and coherence showed relatively lower importance (all below 0.003). This indicates that, among the various factors influencing citation counts, text readability is the most critical linguistic feature.

Figure 2: Feature Importance Ranking Results Based on the XGBoost Model



4.8.3 Logistic Regression

To further reveal the direct influence mechanisms of each feature variable on citation counts (timesCited) and enhance the interpretability of the machine learning results, multiple linear regression models were constructed following the machine learning analysis. This model was selected for its strong parameter interpretability, which directly reveals the marginal contributions and statistical significance of each variable, effectively compensating for the limitations of "black-box" models in mechanistic explanation [65]. The multiple linear regression model is specified as follows:

$$\text{timesCited}=\beta_0+\beta_1\text{readability}+\beta_2\text{emoExpression}+\beta_3\text{lexicalRichness}+\beta_4\text{coherence}+\gamma X+\varepsilon$$

where timesCited is the dependent variable (citation count); readability, emoExpression, lexicalRichness, and coherence are the core explanatory variables (standardized writing style features); X is the vector of all baseline control variables with corresponding coefficient vector γ; and ε is the random error term.

The results of the multiple linear regression analysis are presented in Table 9. The four linguistic style features exhibit significant differences in their effects on citation counts. The readability feature shows a significantly negative coefficient (β = -7.635, p = 0.005), indicating that more complex and less readable text is associated with lower citation counts. The emotional and subjective expression feature has a significantly positive coefficient (β = 7.676, p = 0.010), suggesting that positive emotional expression contributes to greater impact and higher citations. The lexical richness feature exhibits a marginally significant negative effect (β = -8.554, p = 0.051), reflecting that overly complex vocabulary usage may hinder dissemination. The coherence feature also shows a significantly negative association (β = -13.796, p < 0.001), indicating that excessive pursuit of semantic coherence at the sentence level may suppress citations.

Table 9: Regression Analysis Results for Citation Counts (timesCited) and Writing Style Features

Writing style features	DV：timesCited
	Model：OLS
readability	-7.634*** (2.714)
emoExpression	7.676** (2.991)
lexicalRichness	-8.553* (4.375)
coherence	-13.796*** (2.514)
Controls (baseline)	Yes
R-squared	0.280
Adj R-squared	0.279
Number of obs	23,644

*Note: *p < 0.1, **p < 0.05, ***p < 0.01; figures in parentheses are robust standard errors (Std. err.); controls include all baseline features such as number of references, number of authors, funding status, and others.*

5. Discussion and Implications

5.1 Summary of Findings

Through ablation experiments, feature importance analysis, and interpretable modeling, this study systematically explored the mechanisms by which writing style influences paper citation counts.

In the ablation experiments, the results demonstrate that all four categories of writing style features contribute incrementally to model predictive performance. The addition of readability features reduced the model's prediction error (RMSE) by 9.47 units, representing a 7.9% improvement, and exhibited the strongest predictive power. Emotional and subjective expression, lexical richness, and coherence of expression each brought varying degrees of performance gains, confirming the synergistic predictive value of multidimensional writing style features.

Feature importance analysis based on the XGBoost model further revealed the relative contributions of each feature. Among all 29 features, historical usage count (usageCountSince2013) emerged as the most important predictor (importance = 0.383), followed by readability (importance = 0.280), which ranked significantly higher than other linguistic features. This underscores the critical role of text readability in influencing paper dissemination and citations. Although lexical richness, emotional expression, and coherence showed relatively lower importance (all below 0.003), indicating a more indirect or limited role compared to other factors affecting impact, they still provide auxiliary contributions in specific academic contexts—particularly at the micro level of writing, such as moderate emotional tone and textual fluency.

The empirical results from multiple linear regression further clarified the specific mechanisms of each linguistic feature. The regression coefficients indicate significant differences in their effects on citation counts. Readability shows a significantly negative coefficient (β = -7.635, p = 0.005), suggesting that simpler and clearer text is associated with higher citation impact, consistent with findings by McCannon (2019) and Dowling et al. (2018). Emotional and subjective expression exhibits a significantly positive effect (β = 7.676, p = 0.010), indicating that positive emotional expression helps enhance paper impact and attract more citations, supporting Sienkiewicz and Altmann (2016). Lexical richness shows a marginally significant negative effect (β = -8.554, p = 0.051), implying that excessive lexical complexity may hinder dissemination. Coherence also displays a significantly negative association (β = -13.796, p < 0.001), suggesting that overly elaborate semantic connectivity at the sentence level may suppress citations.

5.2 Implications and Limitations

5.2.1 Theoretical Implications

This study offers three main theoretical contributions. First, by adopting a multidimensional interactive perspective, it constructs a novel analytical framework that reveals the synergistic effects of writing style elements on citation behavior, overcoming the limitations of prior research that examined isolated linguistic features. Second, the innovative application of explainable machine learning techniques to scientometrics provides new methodological insights for understanding the complex relationship between writing style and academic impact, particularly in exploring the specific pathways through which different writing dimensions influence citations. Third, the identified differentiated impact patterns—especially the dominant role of readability and its non-linear characteristics—provide empirical evidence that deepens our understanding of the intricate interplay between academic writing and knowledge dissemination, offering valuable reference for refining theories of scholarly communication.

5.2.2 Practical Implications

On the practical side, the findings offer guidance for various academic stakeholders. For researchers, prioritizing improvements in readability—through clearer and more concise expression—should be a primary focus. Incorporating moderate emotional expression while maintaining academic rigor may further enhance a paper's appeal and persuasiveness. For academic journals, incorporating language quality—particularly readability—as an auxiliary evaluation criterion during peer review, and providing corresponding writing guidance to authors, could be beneficial. For research management institutions, organizing academic writing workshops that emphasize readability optimization and expressive techniques may help improve the dissemination effectiveness of research outputs. These evidence-based recommendations could provide useful references for optimizing academic writing practices and promoting the broader dissemination of scientific knowledge.

5.3 Limitations and Future Directions

This study has several limitations. First, although abstracts represent the essence of a paper and their style is crucial for attracting readers, they do not fully equate to the writing style of the full text. Different sections (introduction, methods, discussion, etc.) serve distinct rhetorical functions and may exhibit stylistic variations. Thus, the findings primarily capture the influence of "first-impression" writing style, and caution is needed when generalizing to full-text style. Second, the study primarily reveals statistical associations between writing style features and citation counts rather than strict causal relationships. Future research could adopt experimental designs to establish causality more rigorously. Third, the analysis is based on top-tier journals in three disciplines, ensuring high paper quality but limiting generalizability to lower-tier journals or other fields. Future studies could incorporate a broader range of journals and disciplines to test the universality of the findings.

6. Conclusion

This study analyzed 23,644 papers from computer science, management, and biomedicine. Multidimensional features—including four categories of linguistic indicators—were extracted, and multiple machine learning models combined with grid search hyperparameter optimization were trained and tested. The

XGBoost model ultimately demonstrated the best performance in predicting citation counts. To gain deeper insight into the model's decision-making mechanisms, ablation experiments, feature importance ranking, and regression analysis were employed to systematically investigate the influence mechanisms of each feature on paper impact.

The results confirm that different dimensions of academic writing style possess independent explanatory power for citation counts, a finding that holds even after controlling for a comprehensive set of traditional bibliometric indicators. All four linguistic feature categories were shown to have significant predictive ability for citation counts. Among them, readability emerged as the most important linguistic feature, ranking second only to historical usage count; emotional and subjective expression exhibited a positive promotional effect; while lexical richness and coherence require cautious application, as excessive pursuit of complexity may produce adverse outcomes. These findings reveal the distinct roles played by various dimensions of writing style in academic dissemination and provide empirical evidence for understanding the function of writing style in scholarly communication.

Overall, through a multi-method integrated analytical framework, this study validates the critical role of writing style in academic dissemination and elucidates the differentiated influence mechanisms across writing style dimensions. The research not only offers new empirical evidence for the role of academic writing in knowledge dissemination but also provides concrete guidance for scholars seeking to optimize their writing practices.

References

[1] Cai, L., Tian, J., Liu, J., Bai, X., Lee, I., Kong, X. and Xia, F. Scholarly impact assessment: a survey of citation weighting solutions. Scientometrics. 2019, 118(2), pp. 453-478. https://doi.org/10.1007/s11192-018-2973-6.

[2] Liskiewicz, T., Liskiewicz, G. and Paczesny, J. Factors affecting the citations of papers in tribology journals. Scientometrics. 2021, 126(4), pp. 3321-3336. https://doi.org/10.1007/s11192-021-03870-w.

[3] Wang, M., Zhang, J., Jiao, S. and Zhang, T. Evaluating the impact of citations of articles based on knowledge flow patterns hidden in the citations. PLOS ONE. 2019, 14(11), p. e0225276. https://doi.org/10.1371/journal.pone.0225276.

[4] Aksnes, D. W., Langfeldt, L. and Wouters, P. Citations, citation indicators, and research quality: An overview of basic concepts and theories. Sage Open. 2019, 9(1), p. 2158244019829575. https://doi.org/10.1177/2158244019829575.

[5] McCannon, B. C. Readability and research impact. Economics Letters. 2019, 180, pp. 76-79. https://doi.org/10.1016/j.econlet.2019.02.017.

[6] Sienkiewicz, J. and Altmann, E. G. Impact of lexical and sentiment factors on the popularity of scientific papers. Royal Society Open Science. 2016, 3(6), p. 160140. https://doi.org/10.1098/rsos.160140.

[7] Mammola, S., Piano, E., Doretto, A., Caprio, E. and Chamberlain, D. Measuring the influence of non-scientific features on citations. Scientometrics. 2022, 127(7), pp. 4123-4137. https://doi.org/10.1007/s11192-022-04421-7.

[8] Dowling, M., Hammami, H. and Zreik, O. Easy to read, easy to cite? Economics Letters. 2018, 173, pp. 100-103. https://doi.org/10.1016/j.econlet.2018.09.023.

[9] Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P. and Levitt, J. M. Terms in journal articles associating with high quality: can qualitative research be world-leading? Journal of Documentation. 2023, 79(5), pp. 1110-1123. https://doi.org/10.1108/JD-12-2022-0261.

[10] Giglio, A. D. and Costa, M. U. P. d. The use of artificial intelligence to improve the scientific writing of non-native English speakers. Revista da Associação Médica Brasileira. 2023, 69(9), p. e20230560. https://doi.org/10.1590/1806-9282.20230560.

[11] Vázquez, I. S. Writing with conviction: The use of boosters in modelling persuasion in academic discourses. Revista Alicantina de Estudios Ingleses. 2009, 22, pp. 219-237.

[12] Gonsalves, C., Ludwig, S., de Ruyter, K. and Humphreys, A. Writing for Impact in Service Research. Journal of Service Research. 2021, 24(4), pp. 480-499. https://doi.org/10.1177/10946705211024732.

[13] Martínez, A. and Mammola, S. Specialized terminology reduces the number of citations of scientific papers. Proceedings of the Royal Society B: Biological Sciences. 2021, 288(1948), p. 20202581. https://doi.org/10.1098/rspb.2020.2581.

[14] Ryba, R., Doubleday, Z. A., Dry, M. J., Semmler, C. and Connell, S. D. Better Writing in Scientific Publications Builds Reader Confidence and Understanding. Frontiers in Psychology. 2021, 12, p. 714321. https://doi.org/10.3389/fpsyg.2021.714321.

[15] Amon, J. and Hornik, K. Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals. Journal of Informetrics. 2022, 16(2), p. 101284. https://doi.org/10.1016/j.joi.2022.101284.

[16] Armstrong, J. S. Unintelligible management research and academic prestige. Interfaces. 1980, 10(2), pp. 80-86. https://doi.org/10.1287/inte.10.2.80.

[17] Zhang, Y. and Chen, X. Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval. 2020, 14(1), pp. 1-101. https://doi.org/10.1561/1500000066.

[18] Zhou, W. and Wang, X. Human gene therapy: A scientometric analysis. Biomedicine & Pharmacotherapy. 2021, 138, p. 111510. https://doi.org/10.1016/j.biopha.2021.111510.

[19] Ruan, Q. Z., Chen, A. D., Cohen, J. B., Singhal, D., Lin, S. J. and Lee, B. T. Alternative metrics of scholarly output: The relationship among altmetric score, mendeley reader score, citations, and downloads in plastic and reconstructive surgery. Plastic and Reconstructive Surgery. 2018, 141(3), pp. 801-809. https://doi.org/10.1097/PRS.0000000000004128.

[20] Tonta, Y. and Akbulut, M. Does monetary support increase citation impact of scholarly papers? Scientometrics. 2020, 125(2), pp. 1617-1641. https://doi.org/10.1007/s11192-020-03688-y.

[21] Pagel, P. S. and Hudetz, J. A. Scholarly productivity of united states academic cardiothoracic anesthesiologists: Influence of fellowship accreditation and transesophageal echocardiographic credentials on h-index and other citation bibliometrics. Journal of Cardiothoracic and Vascular Anesthesia. 2011, 25(5), pp. 761-765. https://doi.org/10.1053/j.jvca.2011.03.003.

[22] Annalingam, A., Damayanthi, H., Jayawardena, R. and Ranasinghe, P. Determinants of the citation rate of medical research publications from a developing country. SpringerPlus. 2014, 3(1), p. 140. https://doi.org/10.1186/2193-1801-3-140.

[23] Antoniou, G. A., Antoniou, S. A., Georgakarakos, E. I., Sfyroeras, G. S. and Georgiadis, G. S. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. Annals of Vascular Surgery. 2015, 29(2), pp. 286-292. https://doi.org/10.1016/j.avsg.2014.09.017.

[24] Rostami, F., Mohammadpoorasl, A. and Hajizadeh, M. The effect of characteristics of title on citation rates of articles. Scientometrics. 2014, 98(3), pp. 2007-2010. https://doi.org/10.1007/s11192-013-1118-1.

[25] So, M., Kim, J., Choi, S. and Park, H. W. Factors affecting citation networks in science and technology: focused on non-quality factors. Quality & Quantity. 2015, 49(4), pp. 1513-1530. https://doi.org/10.1007/s11135-014-0110-z.

[26] Baker, H. K., Kumar, S. and Pattnaik, D. Research constituents, intellectual structure, and collaboration pattern in the *Journal of Forecasting*: A bibliometric analysis. Journal of Forecasting. 2021, 40(4), pp. 577-602. https://doi.org/10.1002/for.2731.

[27] Padial, A., Nabout, J., Siqueira, T., Bini, L. and Diniz-Filho, J. Weak evidence for determinants of citation frequency in ecological articles. Scientometrics. 2010, 85(1), pp. 1-12. https://doi.org/10.1007/s11192-010-0231-7.

[28] Amara, N., Landry, R. and Halilem, N. What can university administrators do to increase the publication and citation scores of their faculty members? Scientometrics. 2015, 103(2), pp. 489-530. https://doi.org/10.1007/s11192-015-1537-2.

[29] Yan, Y., Tian, S. and Zhang, J. The impact of a paper's new combinations and new components on its citation. Scientometrics. 2020, 122(2), pp. 895-913. https://doi.org/10.1007/s11192-019-03314-6.

[30] Chinchilla-Rodríguez, Z., Costas, R., Robinson-García, N. and Larivière, V. Examining the quality of the corresponding authorship field in Web of Science and Scopus. Quantitative Science Studies. 2024, 5(1), pp. 76-97. https://doi.org/10.1162/qss_a_00288.

[31] Marx, W. and Bornmann, L. On the causes of subject-specific citation rates in Web of Science. Scientometrics. 2015, 102(2), pp. 1823-1827. https://doi.org/10.1007/s11192-014-1499-9.

[32] Liu, W., Ni, R. and Hu, G. Web of Science Core Collection's coverage expansion: the forgotten Arts & Humanities Citation Index? Scientometrics. 2024, 129(2), pp. 933-955. https://doi.org/10.1007/s11192-023-04917-w.

[33] Chi, P.-S. and Glänzel, W. An empirical investigation of the associations among usage, scientific collaboration and citation impact. Scientometrics. 2017, 112(1), pp. 403-412. https://doi.org/10.1007/s11192-017-2356-4.

[34] Markusova, V., Bogorov, V. and Libkind, A. Usage metrics vs classical metrics: analysis of Russia's research output. Scientometrics. 2018, 114(2), pp. 593-603. https://doi.org/10.1007/s11192-017-2597-2.

[35] Alyousef, H. S. An SF-MDA of the Textual and the Logical Cohesive Devices in a Postgraduate Accounting Course. Sage Open. 2020, 10(3), p. 2158244020947129. https://doi.org/10.1177/2158244020947129.

[36] Lei, L. and Yan, S. Readability and citations in information science: evidence from abstracts and articles of four journals (2003–2012). Scientometrics. 2016, 108(3), pp. 1155-1169. https://doi.org/10.1007/s11192-016-2036-9.

[37] Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M. and Zhang, C. Examining scientific writing styles from the perspective of linguistic complexity. Journal of the Association for Information Science and Technology. 2019, 70(5), pp. 462-475. https://doi.org/10.1002/asi.24126.

[38] de Souza, V. M. A. and Feltrim, V. D. An analysis of textual coherence in academic abstracts written in portuguese. Available from: https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-177.pdf (accessed 26 February 2026).

[39] Rabby, G. and Berka, P. Multi-class classification of COVID-19 documents using machine learning algorithms. Journal of Intelligent Information Systems. 2023, 60(2), pp. 571-591. https://doi.org/10.1007/s10844-022-00768-8.

[40] Liu, J., Wang, X. and Liang, X. Bibliometric feature identification and analysis of retracted papers in biomedicine: An interpretable machine learning perspective. Information Processing & Management. 2025, 62(5), p. 104176. https://doi.org/10.1016/j.ipm.2025.104176.

[41] Alohali, Y. A., Fayed, M. S., Mesallam, T., Abdelsamad, Y., Almuhawas, F. and Hagr, A. A machine learning model to predict citation counts of scientific papers in otology field. BioMed Research International. 2022, 2022(1), p. 2239152. https://doi.org/10.1155/2022/2239152.

[42] Xu, Y., Ju, L., Tong, J., Zhou, C.-M. and Yang, J.-J. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. Scientific Reports. 2020, 10(1), p. 2519. https://doi.org/10.1038/s41598-020-59115-y.

[43] Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K. and Ramachandran, M. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes. 2021, 9(11), p. 2015. https://doi.org/10.3390/pr9112015.

[44] Yan, R., Huang, C., Tang, J., Zhang, Y. and Li, X. To better stand on the shoulder of giants. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, 2012; pp. 51-60.

[45] Wang, M., Jiao, S., Zhang, J., Zhang, X. and Zhu, N. Identification high influential articles by considering the topic characteristics of articles. IEEE Access. 2020, 8, pp. 107887-107899. https://doi.org/10.1109/ACCESS.2020.3001190.

[46] Khan, N., Nauman, M., Almadhor, A. S., Akhtar, N., Alghuried, A. and Alhudhaif, A. Guaranteeing correctness in black-box machine learning: A fusion of explainable AI and formal methods for healthcare decision-making. IEEE Access. 2024, 12, pp. 90299-90316. https://doi.org/10.1109/ACCESS.2024.3420415.

[47] Musolf, A. M., Holzinger, E. R., Malley, J. D. and Bailey-Wilson, J. E. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. Human Genetics. 2022, 141(9), pp. 1515-1528. https://doi.org/10.1007/s00439-021-02402-z.

[48] Żbikowski, K. and Antosiuk, P. A machine learning, bias-free approach for predicting business success using Crunchbase data. Information Processing & Management. 2021, 58(4), p. 102555. https://doi.org/10.1016/j.ipm.2021.102555.

[49] Ferguson, C., Merga, M. and Winn, S. Communications in the time of a pandemic: the readability of documents for public consumption. Australian and New Zealand Journal of Public Health. 2021, 45(2), pp. 116-121. https://doi.org/10.1111/1753-6405.13066.

[50] Basch, C. H., Mohlman, J., Hillyer, G. C. and Garcia, P. Public health communication in time of crisis: Readability of on-line COVID-19 information. Disaster Medicine and Public Health Preparedness. 2020, 14(5), pp. 635-637. https://doi.org/10.1017/dmp.2020.151.

[51] Xu, G., Yu, Z., Yao, H., Li, F., Meng, Y. and Wu, X. Chinese text sentiment analysis based on extended sentiment dictionary. IEEE Access. 2019, 7, pp. 43749-43762. https://doi.org/10.1109/ACCESS.2019.2907772.

[52] Sanders, T. J. and Spooren, W. P. Causality and subjectivity in discourse: The meaning and use of causal connectives in spontaneous conversation, chat interactions and written text. Linguistics. 2015, 53(1), pp. 52-93. https://doi.org/10.1515/ling-2014-0034.

[53] Vögelin, C., Jansen, T., Keller, S. D., Machts, N. and Möller, J. The influence of lexical features on teacher judgements of ESL argumentative essays. Assessing Writing. 2019, 39, pp. 50-63. https://doi.org/10.1016/j.asw.2018.12.003.

[54] Gómez Vera, G., Sotomayor, C., Bedwell, P., Domínguez, A. M. and Jéldrez, E. Analysis of lexical quality and its relation to writing quality for 4th grade, primary school students in Chile. Reading and Writing. 2016, 29(7), pp. 1317-1336. https://doi.org/10.1007/s11145-016-9637-9.

[55] Potratz Jill, R., Gildersleeve-Neumann, C. and Redford Melissa, A. Measurement properties of mean length of utterance in school-age children. Language, Speech, and Hearing Services in Schools. 2022, 53(4), pp. 1088-1100. https://doi.org/10.1044/2022_LSHSS-21-00115.

[56] Shi, Y., Li, Y. and Li, N. Sentence coherence evaluation based on neural network and textual features for official documents. Electronic Research Archive. 2023, 31(6), pp. 3609-3624. https://doi.org/10.3934/era.2023183.

[57] Liang, Z., Mao, J., Lu, K., Ba, Z. and Li, G. Combining deep neural network and bibliometric indicator for emerging research topic prediction. Information Processing & Management. 2021, 58(5), p. 102611. https://doi.org/10.1016/j.ipm.2021.102611.

[58] Forster, E. C. Power and paragraphs: academic writing and emotion. Journal of Learning Development in Higher Education. 2020, 667(19), p. 2020. https://doi.org/10.47408/jldhe.vi19.610.

[59] Ji, S., Sun, W. and Marttinen, P. Content reduction, surprisal and information density estimation for long documents. arXiv preprint arXiv:2309.06009. 2023. https://doi.org/10.48550/arXiv.2309.06009.

[60] Barzilay, R. and Lapata, M. Modeling local coherence: An entity-based approach. Computational Linguistics. 2008, 34(1), pp. 1-34. https://doi.org/10.1162/coli.2008.34.1.1.

[61] Khatoon, A., Daud, A. and Amjad, T. Categorization and correlational analysis of quality factors influencing citation. Artificial Intelligence Review. 2024, 57(3), p. 70. https://doi.org/10.1007/s10462-023-10657-3.

[62] Alhowyan, A., Mahdi, W. A. and Obaidullah, A. J. Computational intelligence investigations on evaluation of salicylic acid solubility in various solvents at different temperatures. Scientific Reports. 2025, 15(1), p. 7142. https://doi.org/10.1038/s41598-025-90704-x.

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

[64] Wang, H., Liang, Q., Hancock, J. T. and Khoshgoftaar, T. M. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. Journal of Big Data. 2024, 11(1), p. 44. https://doi.org/10.1186/s40537-024-00905-w.

[65] Morán-Figueroa, G.-H., Muñoz-Pérez, D.-F., Rivera-Ibarra, J.-L. and Cobos-Lozada, C.-A. Model for predicting maize crop yield on small farms using clusterwise linear regression and GRASP. Mathematics. 2024, 12(21), p. 3356. https://doi.org/10.3390/math12213356.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access