ACTA TECHNICA JAURINENSIS
Vol. XX, No. Y, pp. ZZ-ZZ, 2026
10.14513/actatechjaur.00973

Research Article

Automated Classification of SAP Literature: Predicting Impact and Trends

Kawkab Bouressace[1, *], Tamás Orosz[1]

[1]Faculty of Informatics, Data Science and Engineering Department, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
*e-mail: kawkab@inf.elte.hu

Submitted: 27/01/2026 *Revised: 20/04/2026* *Accepted: 21/04/2026* *Published online: 28/08/2026*

Abstract: The rapid evolution of business systems such as SAP (Systems, Applications, and Products in Data Processing) has generated a growing body of research on implementations, innovations, and business impacts. Determining high-impact papers and detecting emerging trends remains challenging due to the volume of literature. This study presents a machine learning powered pipeline for collecting, preprocessing, and classifying SAP-related research articles retrieved from Semantic Scholar. The pipeline employs natural language processing techniques, including text cleaning, lemmatization, and SciBERT embeddings, to generate richer feature representations, along with metadata features such as vocabulary diversity, novelty score, paper age, and citation velocity. To analyse research impact and trends, we trained a set of classical machine learning models, Random Forest, XGBoost, and LightGBM, and a set of large language models (LLMs), BERT, RoBERTa, and ELECTRA, fine-tuned for classification. The LLMs demonstrated superior performance compared to classical models, achieving accuracies of approximately 93% to 97% for impact classification and 96% to 97% for trend categorization. The models classify papers along two key dimensions: impact classification (High Impact, Niche, Low Impact) and trend categorization (Hot Trend, Recent Classic, Established, Historic), defined using proxy-based bibliometric indicators. This contribution provides an automated framework for literature analysis in the SAP context, enabling researchers and practitioners to identify high-impact studies and support the identification of emerging research directions.

Keywords: SAP; research impact; trend analysis; LLMs; article classification; semantic scholar.

I. INTRODUCTION

Enterprise applications such as SAP are at the core of digital change, business process optimization, and strategic business decision-making across sectors [1]. In engineering-intensive industries, including manufacturing, automotive, logistics, and civil infrastructure, SAP systems serve as the operational backbone for managing production planning, plant maintenance, supply chain coordination, and engineering project lifecycles, making them a critical component of modern industrial IT infrastructure [2-4]. SAP has evolved rapidly over the last decade through tools such as SAP S/4HANA, SAP Fiori, SAP Cloud Platform, and SAP Business Technology Platform (BTP) [5], enabling firms to integrate cloud computing, artificial intelligence (AI), and advanced analytics into their business. These developments are particularly significant in the context of Industry 4.0, where the convergence of cyber-physical systems, intelligent automation, and enterprise software is reshaping engineering operations at scale. This advancement has inspired a growing amount of academic and applied research on SAP implementation, SAP innovation, and business impacts. As literature grows, researchers and practitioners are faced with two persistent challenges, differentiating high-impact research that advances current discourse in enterprise systems, and identifying emerging trends that herald new directions for innovation and adoption [6][7]. Traditional bibliometric methods such as citation counts or classical literature searches, though effective, are likely to fail to deliver insights in a timely fashion due to the excess of publications, lag in citation, and interdisciplinary nature of research topics [8-10]. Moreover, rates of technological change in business systems can cause traditional approaches to be follow-on rather than proactive, thus limiting their usefulness for strategic planning and forecasting. These limitations illustrate a growing requirement for computational, data-driven approaches to the study of literature that apply techniques such as natural language processing (NLP)

[11] and machine learning [12]. Such techniques, in addition to enabling systematic identification of high-impact scholars and works, allow for detection of emerging research frontiers, knowledge clusters, and topics. By providing a more dynamic and scalable perspective of enterprise systems research, these methodologies help bridge the gap between practical decision-making and scholarly research, ultimately influencing both technology adoption and future research agendas. From an information technology engineering perspective, the design and implementation of such an automated pipeline, encompassing data acquisition, feature engineering, embedding-based representation, and supervised classification, constitutes a contribution to the field of intelligent systems engineering applied to knowledge management [13-15]. In this study we focus specifically on the SAP domain, collecting a dataset of research articles using 25 carefully selected SAP-related keywords to ensure coverage of the most relevant literature on SAP tools, technologies, implementation strategies, and business impacts, commonly used by authors in abstracts, titles, or keyword sections. The dataset was sourced from Semantic Scholar, and the collected papers were processed using a machine learning powered pipeline for systematic cleaning, feature extraction, and classification. The pipeline combines NLP steps such as text cleaning and duplicate removal with enriched representations that include SciBERT embeddings [16] and metadata-derived features such as vocabulary diversity, novelty score, paper age, and citation velocity. These enriched features provide a comprehensive representation of research papers beyond surface-level statistics. To analyze research impact and trends, we trained a set of classical machine learning models, Random Forest (RF), XGBoost (XGB), and LightGBM, as well as a set of large language models (LLMs), BERT, RoBERTa, and ELECTRA, by fine-tuning them for classification. The classical models were trained on engineered features and document embeddings, while the LLMs were fine-tuned end-to-end on text representations for direct classification. The models classify papers along two key dimensions, Impact classification (High Impact, Niche, Low Impact) and Trend categorization (Hot Trend, Recent Classic, Established, Historic). These categories are derived from proxy-based bibliometric indicators, enabling supervised classification without manual expert labeling. By bridging bibliometrics, NLP, and both classical and large language model machine learning, this study supports researchers and practitioners in navigating the expanding SAP literature landscape. The proposed framework is especially relevant for engineering organizations and R&D departments seeking to systematically monitor enterprise software research, prioritize technology investments, and align their digital transformation strategies with emerging scholarly directions, as it enables the systematic identification of high-impact research and upcoming topics, thereby enhancing decision-making for academic inquiry, enterprise adoption, and innovation strategies.

The remainder of this paper is organized as follows. Section 2 contains an introduction to related literature in SAP research and machine learning-based approaches for research classification and trend discovery. Section 3 describes the proposed methodology, including data fetching from Semantic Scholar, pre-processing of text, feature engineering, and the use of both classical classifiers (RF, XGB, LightGBM) and large language models (BERT, RoBERTa, ELECTRA). Section 4 presents the results and insights derived from model predictions, along with a discussion of findings regarding trends, research impact, and limitations. Finally, Section 5 concludes the study, summarizing key contributions and directions for future research.

II. Related Work

Existing studies on research impact and trend analysis can be broadly grouped into three categories: bibliometric and network-based analyses, citation and topic growth modeling, and machine learning–based impact classification. This section reviews representative work from each category and highlights their limitations in addressing large-scale, domain-specific literature such as SAP research.

From a bibliometric and network-based perspective, publishing is a fundamental scientific activity, and studies on China–Belt and Road Initiative collaborations show that network analysis and GM (1,1) forecasting can predict high-impact papers, highlighting China's central role, key disciplines, and trends toward diversified, high-quality outputs. However, reliance on historical data and specific forecasting models may not capture sudden shifts in collaboration patterns or emerging research fields [17]. Country-level analyses indicate that citation impact is shaped by multiple factors, including R&D investment, international collaboration, and publication in top-tier journals, with lower-investment countries able to achieve high impact through alternative paths. Nevertheless, macro-level analyses may not capture article-level or field-specific nuances influencing citation outcomes [18]. Trends in a Field of Study (FoS) significantly influence citation counts, with papers within the same FoS exhibiting similar citation patterns and a 66% likelihood of following the same citation trend using the Field of Study Multigraph (FoM) technique and graph centrality measures. However, the analysis is limited to the Computer Science domain and the Microsoft Academic Graph dataset from 2007–2015, which may not generalize across other disciplines, hierarchical levels of FoS, or more recent publication trends [19].

Topic growth and citation dynamics significantly affect citation impact, with publications in fast-growing topics gaining more citations, though this may reflect visibility or hype rather than quality. The analysis is limited to eight disciplines, one publication year, and excludes factors like author prestige and topic novelty, highlighting caution in generalizing the results and in applying citation normalization [20].

Machine learning approaches have been applied for early identification of high-impact papers in fields such as physiology and physics, combining paper, content, and citation features alongside techniques to address imbalanced datasets. While these frameworks achieve strong predictive performance, their reliance on fixed impact thresholds and limited feature sets constrains generalizability [21]. More recent studies propose multi-model frameworks that deploy domain-specific models and incorporate early citation information, achieving improved performance over single-model approaches across heterogeneous research areas. Nonetheless, these methods remain heavily dependent on early citation signals and often exclude broader contextual or semantic features [22].

Beyond modeling approaches, several studies have examined contextual and structural factors affecting citation impact. Research on gender diversity in scientific collaboration reveals that teams with balanced gender representation tend to produce more novel and higher-impact research outputs. An analysis by Yang et al. [23] confirms that gender-diverse teams consistently outperform same-gender teams in terms of novelty and impact, even after controlling for team size, researcher characteristics, and network position. Additional work has evaluated journal-level indicators, such as analyses of Nature Index journals, which highlight strong academic standards and high levels of disruptive innovation across most research topics, while also identifying gaps in underrepresented areas. These findings are limited by reliance on open bibliographic data of varying quality and by the exclusion of humanities and social sciences [24].

Formal characteristics of scientific papers have also been shown to influence citation outcomes. Factors such as title length, article length, and formatting explain a substantial proportion of variance in Field-Weighted Citation Impact, although these results are based on relatively small datasets and require broader validation [25]. Similarly, studies on Open Access (OA) publishing indicate that OA articles receive higher visibility and citation counts compared to subscription-based publications, though these conclusions are often limited to single journals or short timeframes [26].

Recent methodological advances have enriched research analysis through bibliometric mapping, conceptual impact frameworks, and hybrid machine learning architectures. Sahoo et al. [27] effectively track field evolution via keyword co-occurrence networks, yet their metadata-only approach lacks semantic depth. Csernovszky et al. [28] introduce multi-dimensional impact assessment logic, but their framework requires manual instantiation and lacks automated classification capabilities. Similarly, Fazlollahtabar [29] demonstrates the robustness of hybrid ML models for operational decision-making, though these techniques remain unexplored for large-scale literature mining.

Despite these methodological advances, research on citation impact and trend analysis remains fragmented: studies typically focus on narrow disciplines, rely on constrained feature sets, or depend heavily on early citation signals and historical patterns. Crucially, enterprise systems literature lacks dedicated analytical frameworks, and the integration of bibliometric proxies with large language models for scalable, content-aware classification remains unexplored. To address these gaps, this study introduces a unified ML and LLM-driven pipeline tailored to SAP research, fusing deep semantic embeddings with proxy-based bibliometric features to enable automated, dual-axis classification of impact and emerging trends.

III. METHODOLOGY

In this section, we describe our methodology for systematically collecting, preprocessing, extracting features, and classifying SAP-related research articles. The main objective is to provide a data-driven framework for identifying high-impact papers and emerging trends in SAP research as it pertains to engineering enterprises. The proposed methodology integrates bibliometric analysis, natural language processing (NLP), and machine learning techniques to generate actionable insights from SAP-related literature, contributing to information technology engineering through an automated, scalable pipeline for literature intelligence (**Fig. 1**). The proposed pipeline is designed to support the end-to-end classification of scholarly papers into impact and trend categories. It encompasses the full workflow, from automated data acquisition to final category assignment. Built with scalability, reproducibility, and methodological rigor in mind, the pipeline consists of five integrated stages: (1) systematic data collection through API-driven search queries, (2) data cleaning and preprocessing to standardize textual content and metadata, (3) dataset labeling to construct a supervised training corpus, (4) model training using a combination of classical machine learning algorithms and large language models, and

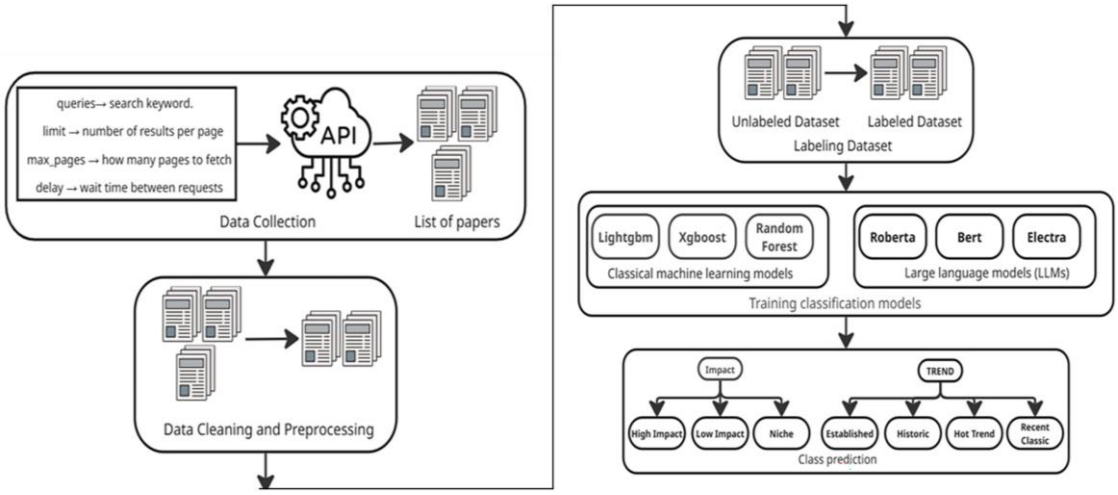

Figure 1. *Automated Pipeline for SAP Literature Classification and Trend classification*

(5) multi-class classification to assign papers to predefined impact and trend categories.

1. Data Collection

The first stage of the proposed pipeline focuses on the systematic collection of SAP-related research articles, forming the foundation for subsequent analysis. Given the breadth and interdisciplinary nature of enterprise systems research, it is crucial to ensure that the dataset is both comprehensive and elevant. For this study, we collected a custom dataset focusing on the SAP domain using targeted queries. A total of 25 keywords (**Table 1**) were used to retrieve academic articles from the Semantic Scholar API, ensuring coverage of a wide range of topics relevant to SAP technologies, modules, and applications.

The dataset was collected programmatically using this API. For each of the 25 keywords, the API was queried with the following parameters: limit=100, which specifies the maximum number of results per page; max_pages=15, defining the number of pages to fetch per keyword; and delay=3, a wait time (in seconds) between requests to avoid rate limiting. For every keyword query, the API returned key metadata, including the title of the article, abstract, publication year, and citation count. The responses were handled according to the status code: a 200 code indicated a successful response, and the data was extracted; a 429 code indicated that the rate limit was been reached, in which case the process paused for 15 seconds and retried; any other response code stopped execution and was logged as an error.

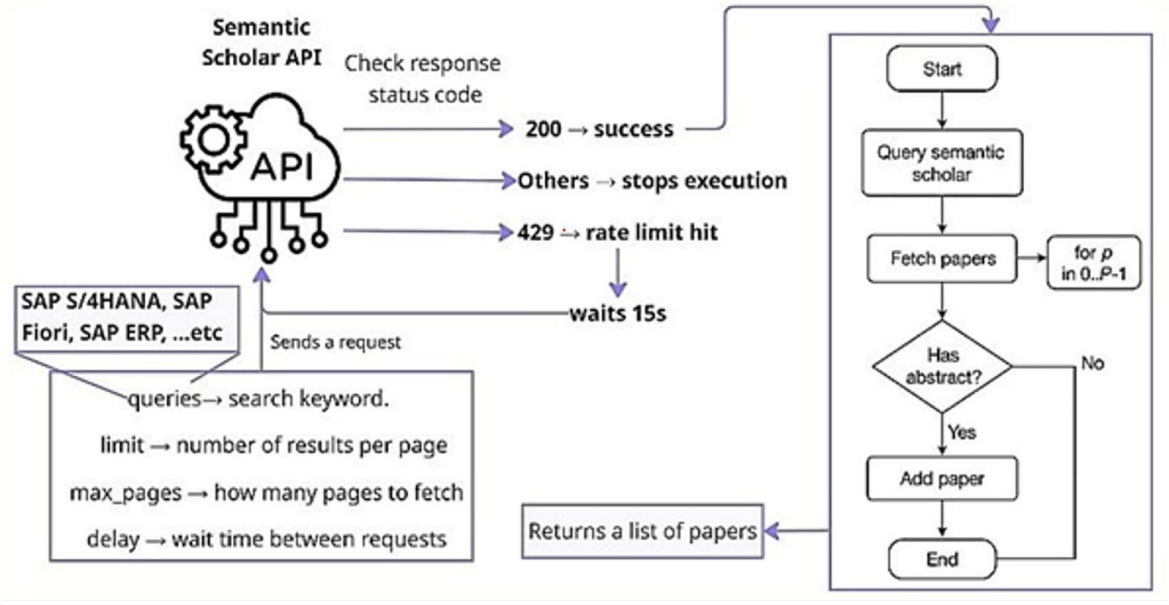

Figure 2. *Overview of Dataset Collection Process*

All articles collected for each keyword were aggregated into a single dataset (see **Fig. 2**), while **Table 1** summarizes the number of pages fetched, papers retrieved, and total papers collected for each keyword.

The dataset consists of 12,900 papers, providing a comprehensive representation of research and development across the SAP domain. The sample in **Table 2** shown here demonstrates the structure of the data, including how titles, abstracts, publication years, and citation counts are organized. This subset allows readers to quickly understand the nature of the dataset without inspecting all entries individually.

This stage prioritizes comprehensive coverage, aiming to collect an extensive list of relevant papers across all facets of SAP research. The resulting dataset serves as the input for the subsequent stages of preprocessing, feature extraction, and machine learning–based classification, where the analysis of research trends and impact will be conducted.

Table 1. *Updated dataset statistics for each search keyword based on actual fetch results*

#	*Search Term*	*Pages Fetched*	*Papers Fetched*	*Total Papers So Far*
1	SAP S/4HANA	1–3, 6, 8–10	700	700
2	SAP Fiori	2, 6, 8–10	500	1200
3	SAP ERP implementation	1–5, 7–8	700	1900
4	SAP business process optimization	1, 4, 7	300	2200
5	SAP cloud platform	1, 2, 4, 6, 8–10	700	2900
6	SAP analytics	1, 2, 9, 10	400	3300
7	SAP supply chain management	1, 3–9	800	4100
8	SAP finance module	2, 3, 5, 6, 8, 9	600	4700
9	SAP SuccessFactors	1–7, 9	600	5500
10	SAP integration	1, 3, 6, 7, 9	500	6000
11	SAP BTP	4–10	700	6700
12	SAP AI core	2, 5, 7	300	7000
13	SAP RISE	2, 3, 5–10	800	7800
14	SAP migration to cloud	–	0	7800
15	SAP data intelligence	3, 5, 6, 9	400	8200
16	SAP customer experience	3, 4, 8, 9	400	8600
17	SAP Ariba	1–3, 5, 8, 10	600	9200
18	SAP Fieldglass	1, 3, 4, 7–9	600	9800
19	SAP Concur	2, 4, 6–8	500	10300
20	SAP sustainability	1, 3, 5, 6, 9	500	10800
21	SAP Leonardo	2–5, 7	500	11300
22	SAP HANA modeling	2, 3, 4, 5, 7, 9	600	11900
23	SAP security and GRC	–	0	11900
24	SAP for industries	1, 3, 5, 7, 9, 10	600	12500
25	SAP intelligent RPA	1, 4–6	300	12900

Table 2. *Sample of SAP Literatures Dataset*

Title	*Abstract*	*Year*	*Citations*
Optimizing supply chain management: strategic ...	In today's highly competitive and rapidly ...	2024	17
Asset Capitalization using SAP S/4HANA in Field ...	In the realm of medical industries, effective ...	2024	10
Unlocking Business Potential: Artificial Intelligence ...	This article explores the transformative role ...	2024	3
Leveraging AI and Machine Learning in SAP S/4HANA ...	This article examines the transformative impact ...	2024	2
Recent Advances and Innovations in SAP S/4HANA ...	This article presents a comprehensive analysis ...	2024	1

2. Data Cleaning and Preprocessing

The dataset was first prepared by carefully cleaning and standardizing the collected research papers. Duplicate records and papers without abstracts were removed to ensure quality and consistency. The text was then pre-processed by converting it to lowercase, removing punctuation and stop words, and applying lemmatization to reduce words to their root forms. To provide a complete representation of each paper's content, the title and abstract were combined into a single text field. This approach captures both the concise summary provided by the title and the detailed overview in the abstract. Combining the title and abstract also ensures a uniform text representation, compensating for variability in detail between papers and providing a consistent basis for analysis. This process created a uniform and informative textual basis for subsequent analysis and modelling. Once the text was ready, the feature engineering phase extracted numerical attributes for every paper. These included the total number of words, the age of the paper (measured as the difference between the current year and its publication year), vocabulary diversity (the ratio of unique words to total words), and a novelty score reflecting the presence of dynamically determined trending terms. Where trending terms were identified based on their frequency within the SAP-specific corpus over the most recent publication years. Specifically, the most frequent words across all papers were identified as trending, and each paper's novelty score was calculated as the proportion of these trending terms present in its text.

Fig. 3 illustrates the progressive reduction of the dataset through each pre-processing stage. Starting with 12,900 initial records, duplicate removal narrowed the pool to 9,205 papers, and the subsequent exclusion of entries with missing abstracts yielded a final cleaned dataset of 3,480 papers. To verify that this final step did not introduce systematic bias, we compared the discarded and retained entries across three dimensions. The abstract-missing papers exhibited a substantially lower median citation count (0 vs. 3) and were heavily concentrated in 2023-2024, indicating they are primarily recent or incompletely indexed records rather than a topically or impact-skewed subset. This conclusion is reinforced by a uniformly distributed retention rate of 55-80% across all 25 keyword queries, with no single topic experiencing disproportionate loss. Together, these patterns confirm that abstract missingness stems from indexing delays rather than content-related bias. Following this validation, a lexical analysis of the cleaned dataset identified the most frequent terms: *sap*, *data*, *system*, *process*, and *business*, effectively mapping the core thematic areas of the SAP literature.

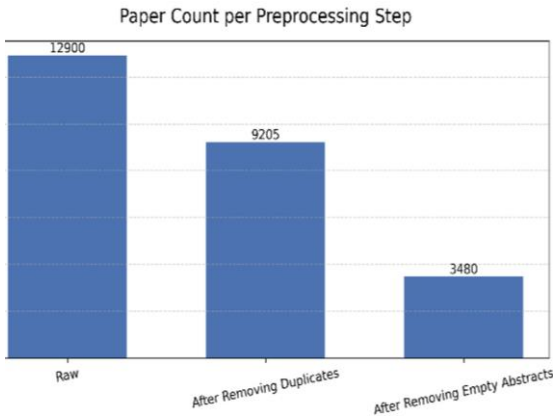

***Figure 3**. Dataset Size Reduction During Preprocessing*

In **Fig. 4**, the original dataset, the distribution of abstract lengths is heavily skewed toward extremely short texts, with a substantial proportion of abstracts containing few or no words, reflecting the presence of incomplete or missing entries. Following pre-processing, which removed duplicates and abstracts that were empty or unusually short, the distribution shifts to the right, peaking around 180–200 words. This indicates that the remaining abstracts are more substantial and informative, yielding a higher-quality dataset. These indicators provide insights into the structural, temporal, and thematic characteristics of each paper, helping to quantify both its scope and relevance. **Table 3** presents a sample of the dataset, showing the organization of both original and newly derived features, allowing readers to quickly understand the dataset without reviewing all entries individually.

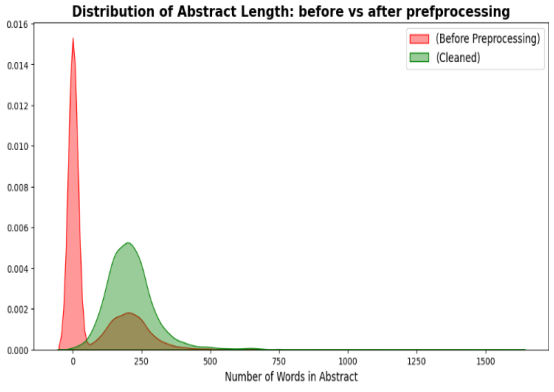

***Figure 4.** Distribution of Abstract Lengths Before and After Preprocessing*

A. Semantic Representations and Feature Matrix

In addition to numerical features, semantic representations were generated using SciBERT embeddings, which are specifically designed for

scientific text. Unlike simple keyword-based metrics, embeddings capture the contextual meaning of words and phrases, enabling a deeper understanding of nuanced similarities and thematic overlaps between papers. Each paper's combined title and abstract were transformed into a dense vector representation, encoding semantic information about its research content. These embeddings were then merged with the numerical features to form a comprehensive feature matrix that balanced both textual semantics and quantitative attributes. To ensure robustness, numerical features were standardized so that variables with different scales, such as word count and paper age, could be compared fairly.

Table 3. *Sample of SAP Literatures Dataset After Preprocessing*

Title	*Abstract*	*Year*	*Citations*	*Text*	*Num Words*	*Paper Age*
Optimizing supply chain management: strategic ...	In today's highly competitive and rapidly evolving ...	2024	17	optimizing supply chain management strategic business ...	265	1
Asset Capitalization using SAP S/4HANA in Field ...	In the realm of medical industries, effective ...	2024	10	asset capitalization using sap hana field inventory ...	139	1
Unlocking Business Potential: Artificial Intelligence ...	This article explores the transformative role ...	2024	3	unlocking business potential artificial intelligence ...	252	1
Leveraging AI and Machine Learning in SAP S/4HANA ...	This article examines the transformative impact ...	2024	2	leveraging ai machine learning sap hana cloud ...	151	1
Recent Advances and Innovations in SAP S/4HANA ...	This article presents a comprehensive analysis ...	2024	1	recent advance innovation sap hana cloud sap business ...	147	1

Missing values in the numerical features, including word count, paper age, vocabulary diversity, and novelty score, were handled using K-Nearest Neighbours (KNN) imputation, which estimates missing entries based on the values of the nearest neighbours in the feature space. This approach prevented bias or data loss and ensured that every paper had a complete feature vector. The resulting enriched and balanced feature set provided a strong foundation for later tasks such as classification, trend detection, and predictive modelling (**Fig. 5**).

Figure 5*. Generating Dense Text Embeddings Using SciBERT*

3. Labeling Papers

To analyse scholarly contributions in the SAP research domain, each paper is assigned two complementary labels: Impact and Trend. These labels are derived from a set of quantitative metrics capturing recognition, novelty, and temporal relevance (see **Table 4**). Thresholds for Impact and Trend labelling were defined using rules computed over the full dataset, ensuring robustness across different publication years and citation scales. This approach avoids manual expert labelling while maintaining consistency and reproducibility.

A. Impact Label

Papers are labelled as High Impact if they exhibit strong recognition, reflected by high citation counts or citation velocity, in combination with above-average vocabulary diversity. Papers that do not meet these criteria are labelled Low Impact, indicating comparatively limited scholarly influence. Formally, this classification can be expressed as:

$$L_{impact} = \begin{cases} High\ Impact, (C > 10\ or\ V > 3) and\ D > 0.2 \\ Low\ Impact, otherwise \end{cases} \quad (1)$$

B. Trend label

Papers are labelled as Hot Trend if they are recent and exhibit rapid citation growth or high novelty, indicating emerging topics. Publications classified as recent classic are fairly new, with moderate novelty and stable influence.

Papers from the last decade that have maintained relevance are labelled Established, while older foundational works with lower ongoing momentum are labelled Historic.

Formally, this classification is expressed as:

$$L_{Trend} = \begin{cases} Hot\ Trend, \\ if\ year \geq Y_{max} - 3\ and\ (V > 10\ or\ N > 0.7) \\ Recent\ Classic, if\ year \geq Y_{max} - 5\ and\ D > 0.65 \\ and \\ V \leq 10\ and\ N \leq 0.7 \\ Established, if\ Y_{max} - 10 \leq year < Y_{max} - 5 \\ Historic, otherwise \end{cases} \quad (2)$$

Trend categories are mutually exclusive and jointly exhaustive, with assignment determined by the combined evaluation of paper age, citation velocity, and novelty score.

C. Label Distributions

Table 5 provides a comprehensive overview of the number of papers in each category, illustrating the relative representation of different levels of impact and research trends within the dataset, which the label distributions exhibit moderate class imbalance, particularly in the 'Established' trend group. This skew may bias classifiers toward majority classes and reduce sensitivity for minority categories, necessitating mitigation strategies during model training.

Table 4. *Impact and Trend Metrics for Paper Classification*

Metric	*Equation*	*Description*
Temporal Metric		
Paper Age (A)	A = Current Year − Publication Year	Age of the paper in years; used to compute Citation Velocity and Trend Label.
Citation Metrics		
Citation Count (C)	C = Total citations of the paper	Total citations received by the paper. Used for Impact Label.
Citation Velocity (V)	V = C / A	Average citations per year; used in both Impact and Trend Labels.
Text Metrics		
Vocabulary Diversity (D)	D = |W_unique| / (|W_total| + 1)	Ratio of unique words to total words, measures conceptual richness. Used in both labels (The +1 in the denominator prevents division by zero for empty texts and slightly smooths scores for very short texts).
Novelty Score (N)	N = (1/T) Σ (i=1 to T) 1{t_i ∈ paper text}	Fraction of trending terms present in text. Used in both labels.

Table 5. *Impact and Trend Label Counts*

Label Type	*Label*	*Count*
Impact	High Impact	1973
	Low Impact	1507
Trend	Historic	1214
	Established	912
	Recent Classic	905
	Hot Trend	449

4. Model Training and Evaluation

To address both impact and trend prediction tasks, we employed a set of classical machine learning algorithms and modern large language model (LLM)-based hybrid architectures. For classical machine learning, we trained LightGBM, XGBoost, and Random Forest classifiers separately for each task. While the labelling rules employ fixed thresholds, classical ensemble algorithms are specifically leveraged for their capacity to model non-linear feature interactions and adaptive decision boundaries. This ensures that predictions generalize beyond simple rule-matching, enabling robust classification of papers with nuanced or borderline bibliometric profiles. These models were selected due to their strong performance on heterogeneous tabular data and their robustness in classification tasks combining textual and bibliometric features. Hyperparameters such as the number of estimators, learning rate, and maximum depth were optimized independently for each task. The reported configurations correspond to the best-performing settings obtained after multiple experimental trials and validation runs conducted in this study, rather than values adopted directly from prior work. To handle class imbalance, we used a two-stage protocol: models were first trained with balanced class weights, and SMOTE was applied only if minority-class recall fell below 0.65, a threshold selected empirically to balance sensitivity with overall performance. For Random Forest, weights were clipped to [0.1, 10] to prevent extreme values from destabilizing individual trees, while LightGBM and XGBoost used unclipped weights due to their built-in regularization mechanisms that naturally handle weight extremes. The final strategy was chosen based on the higher macro-averaged F1-score on the validation set. LightGBM and XGBoost met the recall threshold with weights alone, making SMOTE unnecessary. Random Forest, however, required SMOTE to recover minority recall (Impact: 0.22 to 0.58; Trend: 0.17 to 0.29) and improve macro-F1, demonstrating that adaptive imbalance handling tailored to each model's behaviour yields optimal results (**Table 6**).

Table 6. *Class Imbalance Handling Strategy*

Model	*Task*	*Minority Class*	*Initial Strategy*	*Recall (Weights)*	*Recall (SMOTE)*	*Macro-F1 (Weights)*	*Macro-F1 (SMOTE)*	*Final Strategy*
LightGBM	Impact	Low Impact	Balanced class weights	0.66	—	0.786	—	Balanced class weights
	Trend	Hot Trend	Balanced class weights	0.61	0.63	0.893	0.735	Balanced class weights
XGBoost	Impact	Low Impact	Balanced class weights	0.73	—	0.784	—	Balanced class weights
	Trend	Hot Trend	Balanced class weights	0.77	—	0.922	—	Balanced class weights
Random Forest	Impact	Low Impact	Balanced class weights (clipped)	0.22	0.58	0.711	0.740	SMOTE
	Trend	Hot Trend	Balanced class weights (clipped)	0.17	0.29	0.655	0.705	SMOTE

For LLM-based models, we implemented BERT- and RoBERTa-based hybrid classifiers that combine text embeddings from transformer backbones with numerical features. Each model was trained separately for impact and trend classification to avoid task interference, using focal loss with class weights to address class imbalance. The final hybrid classifier can be represented as:

$$y = f([Embeddings(x_{text}), x_{numerical}]) \quad (3)$$

where Embeddings(x_{text}) are the transformer-based text embeddings and $x_{numerical}$ are the bibliometric features.

Table 7 and **Table 8** summarize the configurations and training setups for each target and model type. For the impact and trend classification tasks, we employed large language models within a hybrid architecture that integrates transformer-based textual representations with structured numerical features. Separate models were trained for each task using BERT, RoBERTa, or ELECTRA backbones to avoid interference between impact and trend prediction. Textual inputs were encoded using pretrained transformer models, with the resulting embeddings projected into a 256-dimensional latent space. Numerical features, including vocabulary diversity, novelty score, paper age, and citation counts, were linearly projected into a 64-dimensional representation. These embeddings were concatenated and processed by a fusion layer implemented as a feed-forward network, enabling joint modeling of semantic and bibliometric information, before the final classification layer produced the predicted labels.

Table 7. *Classic Machine Learning Models for Impact and Trend classification*

Model	*Target*	*Hyperparameters / Key Settings*	*Class Imbalance Handling*
LightGBM	Impact	n_estimators=1000, lr=0.1, max_depth=50	Optional SMOTE
	Trend	n_estimators=500, lr=0.01, max_depth=8	Optional SMOTE
XGBoost	Impact	n_estimators=1000, lr=0.01, max_depth=6, subsample=0.8, colsample_bytree=0.7	Sample weights
	Trend	n_estimators=1000, lr=0.01, max_depth=6, subsample=0.8, colsample_bytree=0.7	Sample weights
Random Forest	Impact	n_estimators=300–1000, max_depth=20–60, min_samples_split=2–10, max_features={sqrt, log2}	Balanced & clipped class weights
	Trend	n_estimators=400–1000, max_depth=25–60, min_samples_split=2–5, max_features=sqrt	Balanced & clipped class weights

Table 8. *Dataset split distribution*

Joint Stratum (Impact × Trend)	*Full Dataset (N)*	*Train 80% (n=2,784)*	*Validation 10% (n=348)*	*Test 10% (n=348)*
High Impact × Historic	688	550	69	69
High Impact × Established	517	414	52	51
High Impact × Recent Classic	513	410	52	51
High Impact × Hot Trend	255	204	26	25
Low Impact × Historic	526	421	53	52
Low Impact × Established	395	316	40	39
Low Impact × Recent Classic	392	314	39	39
Low Impact × Hot Trend	194	155	20	19
Total	3,480	2,784	348	348

To mitigate class imbalance, focal loss with class-specific weighting was applied, placing greater emphasis on underrepresented categories. The cleaned dataset (N=3,480) was partitioned into training (80%, n=2,784), validation (10%, n=348), and test (10%, n=348) sets using stratified sampling on the joint label formed by combining Impact and Trend categories (e.g., HighImpact_HotTrend). Joint stratification was applied uniformly across all eight strata because each contained sufficient instances to ensure proportional representation in every split under the 80/10/10 configuration. All splits used a fixed random seed (random_state=42) to ensure reproducibility. **Table 8** presents the exact distribution of instances across all joint strata for each dataset split.

Model training and evaluation were conducted using the Hugging Face Trainer framework with configurable hyperparameters, including batch size, learning rate, number of epochs, weight decay, gradient clipping, and mixed-precision training. Final evaluations were performed on unseen test data using standard classification metrics, and trained models, tokenizers, and feature definitions were retained to ensure reproducibility. Overall, this hybrid LLM-based approach enables effective integration of semantic textual information with numerical indicators, contributing to improved performance across both impact and trend classification tasks (**Fig. 6** and **Table 9**).

Table 9. *Large Language Model (LLM) Hybrid Classifiers for Impact and Trend*

Model	*Target*	*Hyperparameters / Key Settings*	*Class Imbalance Handling*
RoBERTa / BERT / ELECTRA	Impact	**Transformer backbone: RoBERTa-base / BERT-base-uncased / google/electra-base-discriminator** Text embedding projection: 256 Numerical embedding projection: 64 Fusion layer: 256 Epochs: 11 Batch size: 4 Learning rate: 2e-5	Focal loss with class weights
RoBERTa / BERT / ELECTRA	Trend	**Transformer backbone: RoBERTa-base / BERT-base-uncased / google/electra-base-discriminator** Text embedding projection: 256 Numerical embedding projection: 64 Fusion layer: 256 Epochs: 30 Batch size: 8 Learning rate: 1e-2	Focal loss with class weights

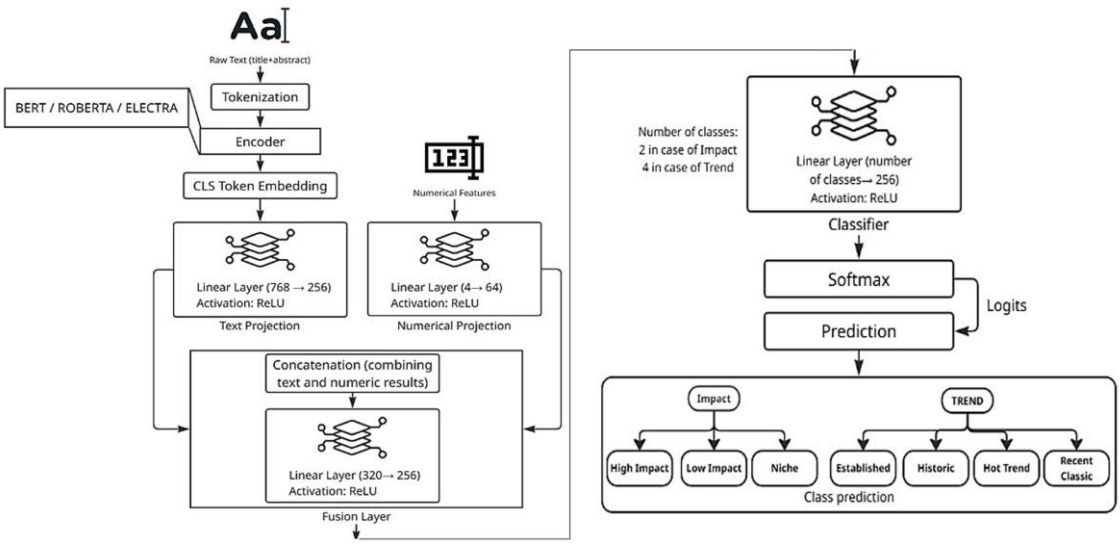

Figure 6. *Workflow for Impact & Trend Prediction Using LLMs and Tabular Feature*

IV. RESULTS AND DISCUSSIONS

This section presents and analyses the experimental results obtained from the proposed impact and trend classification framework. The performance of classical machine learning models and large language models is evaluated and compared across both tasks using standard classification metrics.

We evaluated six models on two classification tasks: Impact Classification (binary: High vs. Low Impact) and Trend Classification (four classes: Established, Historic, Hot Trend, Recent Classic). The models were categorized into Classic Machine Learning methods (LightGBM, XGBoost, and Random Forest) and Large Language Models (RoBERTa-base, BERT-base-uncased, and ELECTRA-base-discriminator). Their performance metrics are reported in **Table 8** and **Table 10**.

Large language models show a clear advantage in Impact Classification. ELECTRA-base-discriminator achieves the highest accuracy (97.13%) with balanced F1 scores (0.974 for High Impact, 0.967 for Low Impact), while BERT and RoBERTa also perform strongly (96.12% and 93.39% accuracy). Classic models lag significantly: LightGBM, the best among them, reaches only 79.74% accuracy and misclassifies one-third of Low Impact samples (recall = 66.2%). Random Forest performs worst, with just 57.6% recall for Low Impact. These results indicate that tree-based models struggle to capture the subtle linguistic cues that determine impact severity (**Table 9**).

Table 10. *Impact Classification Performance across Models*

Model Type	*Model*	*Class*	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score*
Classic ML	LightGBM	High Impact	79.74	0.777	0.901	0.834
		Low Impact		0.837	0.662	0.739
	XGBoost	High Impact	78.59	0.778	0.835	0.806
		Low Impact		0.796	0.730	0.762
	Random Forest	High Impact	75.72	0.734	0.896	0.807
		Low Impact		0.809	0.576	0.673
LLMs	RoBERTa-base	High Impact	93.39	0.946	0.937	0.941
		Low Impact		0.918	0.931	0.924
	BERT-base-uncased	High Impact	96.12	0.974	0.957	0.965
		Low Impact		0.945	0.967	0.956
	ELECTRA-base	High Impact	97.13	0.985	0.965	0.974
		Low Impact		0.955	0.980	0.967

This gap is further highlighted in **Fig. 7** which shows ELECTRA maintaining high precision (0.985) at high recall (0.965), while LightGBM achieves much lower precision (0.777) even at slightly lower recall (0.901), reflecting its bias toward the High Impact class.

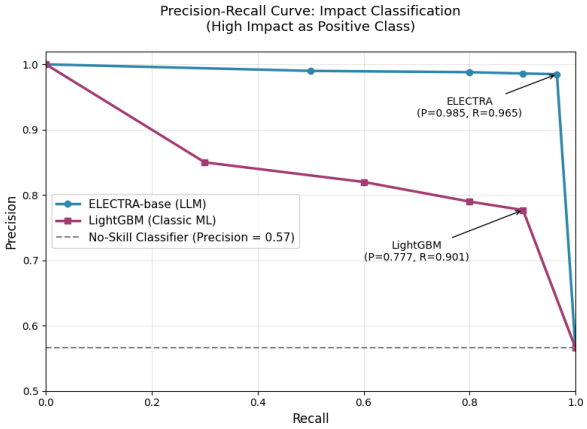

Figure 7. *Precision-Recall curve for Impact Classification (High Impact as positive class)*

The performance gap between large language models and classic approaches is most evident in the Trend Classification task. BERT base uncased achieves the highest accuracy at 97.70 percent and excels on the challenging Hot Trend class, with a recall of 0.989 and an F1 score of 0.983. In contrast, even the best classic model, XGBoost (93.82 percent accuracy), struggles with emerging trends, recalling only 76.7 percent of Hot Trend instances.

This limitation arises because classic models lack contextualized Language representations and therefore cannot generalize to novel or rapidly evolving semantic patterns. The issue is clearly visible in the confusion matrices.

Fig. 8 shows XGBoost misclassifying 21 Hot Trend samples as Historic or Recent Classic, while reveals that BERT makes only one error across all 696 samples. As shown in **Fig. 9,** the BERT confusion

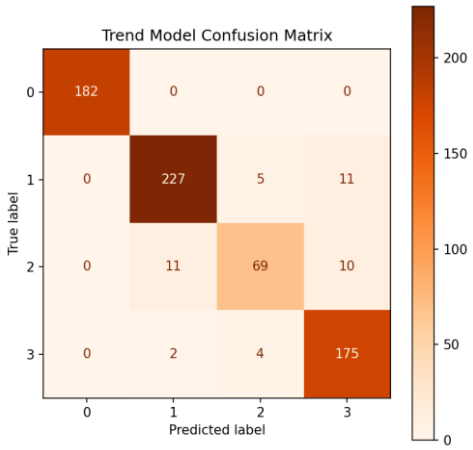

Figure 8. *Confusion matrix XGBoost*

matrix confirms this near-perfect classification of Hot Trend papers, with only a single sample misassigned to another class.

These results align with the quantitative metrics in **Table 11**, which report a macro F1 score of 0.922 for XGBoost compared to 0.978 for BERT. Together, these findings demonstrate that large language models leverage deep contextual understanding to accurately identify dynamic and novel patterns, whereas classic models, constrained by non-contextual features, consistently confuse emerging trends with established categories.

Large language models (LLMs) significantly outperform classic machine learning methods in both Trend Classification and Impact Classification tasks, with accuracy advantages ranging from nearly 4 to over 21 percentage points. LLMs maintain high precision and recall across all classes, including minority and semantically volatile categories such as Low Impact and Hot Trend. This robustness stems from their pretrained contextual representations, which encode deep linguistic knowledge and enable fine-grained semantic discrimination. In contrast, classical models are trained on the same numerical features (citation velocity, vocabulary diversity, novelty score) used to construct the proxy labels. Consequently, their high accuracy primarily reflects the recovery of the underlying rule structure rather than independent generalization. These results therefore quantify the internal consistency of the labeling framework, rather than establishing external predictive validity against independently assessed scholarly impact. They tend to favour dominant classes and miss emerging patterns, limiting their effectiveness in real-world scenarios where fairness, adaptability, and consistent accuracy across diverse content types are essential. Therefore, for applications involving impact assessment or trend

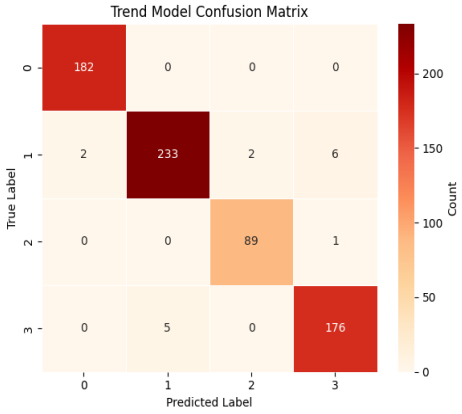

Figure 9. *Confusion matrix BERT*

detection in textual data, large language models represent a significantly more effective and reliable solution.

Table 11. *Trend Classification Performance Across Models*

Model Type	*Model*	*Class*	*Accuracy (%)*	*Precision*	*Recall*	*F1-Score*
Classic ML	LightGBM	Established	91.95	1.000	1.000	1.000
		Historic		0.934	0.934	0.934
		Hot Trend		0.965	0.611	0.748
		Recent Classic		0.822	0.972	0.891
	XGBoost	Established	93.82	1.000	1.000	1.000
		Historic		0.946	0.934	0.940
		Hot Trend		0.885	0.767	0.821
		Recent Classic		0.893	0.967	0.928
	Random Forest	Established	76.15	0.868	0.830	0.848
		Historic		0.711	0.819	0.761
		Hot Trend		0.722	0.289	0.413
		Recent Classic		0.748	0.851	0.796
LLMs	RoBERTa-base	Established	96.41	0.989	0.995	0.992
		Historic		0.978	0.926	0.951
		Hot Trend		0.957	0.989	0.973
		Recent Classic		0.926	0.972	0.949
	BERT-base-uncased	Established	97.70	0.989	1.000	0.995
		Historic		0.979	0.959	0.969
		Hot Trend		0.978	0.989	0.983
		Recent Classic		0.962	0.972	0.967
	ELECTRA-base	Established	96.84	0.989	1.000	0.995
		Historic		0.979	0.938	0.958
		Hot Trend		0.967	0.989	0.978
		Recent Classic		0.936	0.967	0.951

Fig. 10 presents a heatmap of per-class F1-Scores for the Trend Classification task, with trend categories on the horizontal axis and models on the vertical axis.

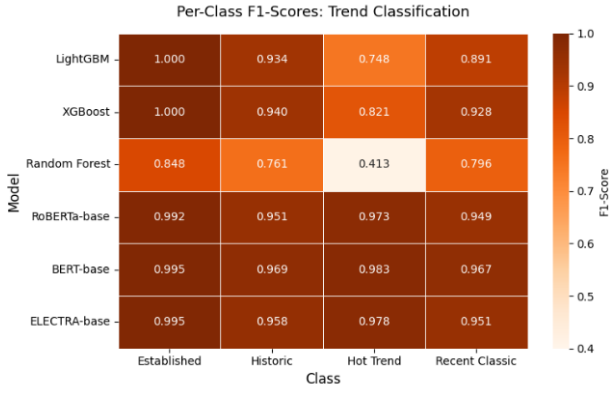

Figure 10. *Pre-class F1-Scores: Trend Classification*

Lighter orange indicates lower performance, while darker orange reflects higher F1-Scores. The plot reveals that classic models, especially Random Forest (F1 = 0.413 on Hot Trend), struggle with emerging categories, whereas large language models maintain consistently high scores across all classes. BERT-base achieves the best performance on Hot Trend with an F1-Score of 0.983.

Fig. 11 shows the same for Impact Classification, comparing High Impact and Low Impact. Classic models exhibit imbalance, with notably lower F1-Scores on Low Impact (e.g., 0.673 for Random Forest).

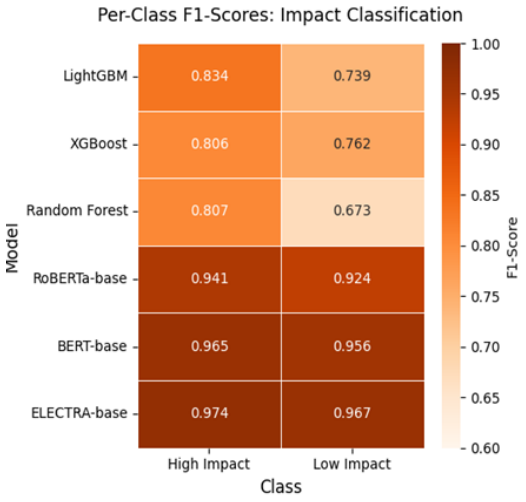

Figure 11. *Pre-class F1-Scores: Impact Classification*

In contrast, large language models show balanced high performance, and ELECTRA-base achieves F1-Scores of 0.974 and 0.967 for High and Low Impact respectively. These results confirm that large language models provide more reliable and equitable classification across all categories.

Limitations of proxy-based supervision and result variability. This study is subject to two methodological constraints that warrant consideration when interpreting the reported results. all experiments utilized a fixed random seed (42) for data partitioning and model initialization to ensure reproducibility. Given the substantial computational cost of fine-tuning large language models, training was not repeated across multiple seeds; consequently, the reported metrics represent point estimates without confidence intervals. Nevertheless, the pronounced performance gap between classical approaches (~76–80% accuracy) and LLM-based classifiers (93–97%) is sufficiently large to indicate that seed-induced variability is unlikely to alter the primary conclusions regarding architectural superiority. A definitive resolution of both limitations requires future validation against expert-annotated gold standards or temporally stratified benchmarks, alongside multi-seed retraining to quantify result stability.

V. Conclusion

This study demonstrates the effectiveness of integrating machine learning and natural language processing techniques to automate the analysis of SAP-related research literature. By developing a pipeline that leverages SciBERT embeddings, metadata-driven features, and advanced classifiers, we implement and evaluate an applied framework for identifying high-impact papers and detecting emerging research trends in the SAP domain; the individual components are established in the literature, and the contribution lies in their systematic integration, domain-specific operationalisation, and comparative evaluation. The results indicate that large language models such as BERT, RoBERTa, and ELECTRA significantly outperform classical machine learning approaches including Random Forest, XGBoost, and LightGBM in both impact and trend classification tasks, achieving accuracies exceeding 93%. The findings illustrate how NLP and deep learning can be combined with bibliometric indicators to support automated literature analysis. This approach enhances the visibility of influential works and provides a foundation for proactive research planning, helping scholars and practitioners navigate the rapidly expanding SAP ecosystem. Furthermore, the integration of automated trend detection within bibliometric workflows enables institutions and enterprises to align research and innovation strategies with evolving technological directions in enterprise systems. Incorporating mechanisms for continuous data updates, adaptive learning, and enhanced citation modeling could further improve the system's responsiveness, scalability, and interpretability, supporting even more effective literature analysis in dynamic research environments.

VI. Acknowledgment

We sincerely thank the Stipendium Hungaricum Scholarship, funded by the Hungarian Government through the Tempus Public Foundation, for making this academic pursuit possible. We also gratefully acknowledge the Department of Data Science at Eötvös Loránd University (ELTE) for their invaluable academic guidance and technical support throughout this research.

VII. Author Contributions

K. Bouressace: Conceptualization, Methodology, Resource provision, Data curation, Analytical review, Visualization, Software, and Writing.

T. Orosz: Supervision, Methodology, Validation, and Writing, review & editing.

VIII. Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jampani, S., Mokkapati, C., Chinta, U., Singh, N., Goel, O., Chhapola, A.: Application of AI in SAP implementation projects. International Journal of Applied Mathematics & Statistical Sciences (IJAMSS) 11(2), pp. 327–350 (2022).

[2] Maha, A., et al.: Optimizing supply chain management: strategic business models and solutions using SAP S/4HANA. Magna Scientia Advanced Research and Reviews 11(01), pp. 339–351 (2024).

[3] Chenna, K.: Integrated SAP S/4HANA production planning with AGI warehouse management system and laboratory information management system. International Journal of Advanced Manufacturing Technology 143, pp. 2393–2404 (2026). https://doi.org/10.1007/s00170-026-17542-7

[4] Shah, V., Shah, S.P.: ERP-Integrated Inventory Control and Profitability in Construction: A Case Study. Journal of Information Systems Engineering and Management 10 (40s), 437 (2025). https://doi.org/10.52783/jisem.v10i40s.7314

[5] Rahman, M.A., Bhowmik, J., Ahamed, M.S., Rahman, R.: Opportunities and challenges in data analysis using SAP: A review of ERP software performance. International Journal of Management Information Systems and Data Science 1(4), pp. 50–67 (2024). https://doi.org/10.62304/ijmisds.v1i04.192

[6] Jaramillo-Mediavilla, L., Basantes-Andrade, A., Cabezas-González, M., Casillas-Martín, S.: Impact of gamification on motivation and academic performance: A systematic review. Education Sciences 14(6), 639 (2024). https://doi.org/10.3390/educsci14060639

[7] Zhou, S., Campbell, L.N., Fyffe, S.: Quantifying the scientist–practitioner gap: How do small business owners react to our academic articles? Industrial and Organizational Psychology 17(4), pp. 379–398 (2024). https://doi.org/10.1017/iop.2024.11

[8] Öztürk, O., Kocaman, R., Kanbach, D.K.: How to design bibliometric research: An overview and a framework proposal. Review of Managerial Science 18(11), pp. 3333–3361 (2024). https://doi.org/10.1007/s11846-024-00738-0

[9] Lazarides, M.K., Lazaridou, I.-Z., Papanas, N.: Bibliometric analysis: Bridging informatics with science. The International Journal of Lower Extremity Wounds 24(3), pp. 515–517 (2025). https://doi.org/10.1177/15347346231153538

[10] Ninkov, A., Frank, J.R., Maggio, L.A.: Bibliometrics: Methods for studying academic publishing. Perspectives on Medical Education 11(3), pp. 173–176 (2022). https://doi.org/10.1007/s40037-021-00695-4

[11] Patwardhan, N., Marrone, S., Sansone, C.: Transformers in the real world: A survey on NLP applications. Information 14(4), 242 (2023). https://doi.org/10.3390/info14040242

[12] Sharifani, K., Amini, M.: Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal* 10(07), pp. 3897–3904 (2023)

[13] Gao, Y., Chen, L., Wang, X., et al.: CTXPipe: Context-aware automated data pipeline design for machine learning. IEEE Transactions on Knowledge and Data Engineering 35(6), pp. 7499–7519 (2024)

[14] Elkholy, A., Hassan, M., Al-Masri, E.: Interpretable automated machine learning in enterprise knowledge management systems. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1937–1945 (2024)

[15] Zhang, B., Wu, Y., Fan, R., et al.: Automated feature engineering for knowledge graph-enhanced document classification. Knowledge-Based Systems 321, 113671 (2025)

[16] Maheshwari, H., Singh, B., Varma, V.: SciBERT sentence representation for citation context classification. In: Proceedings of the Second Workshop on Scholarly Document Processing, pp. 130–133 (2021). https://aclanthology.org/2021.sdp-1.17/

[17] Wei, M., Savage, R.: Leadership of Scientific Cooperation and Forecasting: High-Impact Papers on Scientific Collaboration between China and the Belt and Road Countries. SSRN 4826880 (2025). URL: https://ssrn.com/abstract=4826880

[18] McManus, C., Neves, A.A.B., Diniz Filho, J.A., Pimentel, F., Pimentel, D.: Funding as a determinant of citation impact in scientific papers in different countries. Anais da Academia Brasileira de Ciências 95(1), 20220515 (2023)

[19] Zafar, L., Masood, N.: Impact of field of study trend on scientific articles. IEEE Access 8, pp. 128295–128307 (2020). https://doi.org/10.1109/ACCESS.2020.3007558

[20] Sjögårde, P., Didegah, F.: The association between topic growth and citation impact of research publications. Scientometrics 127(4), 1903–1921 (2022). https://doi.org/10.1007/s11192-022-04293-x

[21] Hu, Z., Cui, J., Gu, Y., Qammar, R.: Multi-dimensional indicator system construction and prediction of potentially high-impact papers combining machine learning models and unbalanced sampling. *Authorea Preprints* (2025)

[22] Zhang, F., Wu, S.: Predicting citation impact of academic papers across research areas using multiple models and early citations. Scientometrics 129(7), pp. 4137–4166 (2024). https://doi.org/10.1007/s11192-024-05086-0

[23] Yang, Y., Tian, T.Y., Woodruff, T.K., et al.: Gender-diverse teams produce more novel and higher-impact scientific ideas. Proceedings of the National Academy of Sciences 119 (36), 2200841119 (2022). https://doi.org/10.1073/pnas.2200841119

[24] Jiang, Y., Ma, J.: Are Nature Index journals a valid basis for academic assessment: A study of academic impact and disruptive innovation assessment based on open bibliographic metadata and citation data. Humanities and Social Sciences Communications 12, article 1062 (2025). https://doi.org/10.1057/s41599-025-05387-6

[25] Dengler, J.: Determinants of citation impact. Vegetation Classification and Survey 5, pp. 169–177 (2024). https://doi.org/10.3897/VCS.126956

[26] Yi, H., Cao, Y., Leng, Q., Wang, Y., Zhang, G., Mao, Y.: The impact of open access on citations, pageviews, and downloads: A scientometric analysis in Postgraduate Medical Journal. Postgraduate Medical Journal 100 (1187), pp. 679–685 (2024). https://doi.org/10.1093/postmj/qgae047

[27] Sahoo, S.K., Choudhury, B.B., Dhal, P.R.: A Bibliometric Analysis of Material Selection Using MCDM Methods: Trends and Insights. Spectrum of Mechanical Engineering and Operational Research 1(1), pp. 189–205 (2024). https://doi.org/10.31181/smeor11202417

[28] Csernovszky, A., Szalmane Csete, M.: Artificial Intelligence and Sustainability: A Conceptual Framework for System-Level Impact Assessment. Cognitive Sustainability 5(1) (2026). https://ojs.mtak.hu/index.php/CogSust/article/view/22409

[29] Fazlollahtabar, H.: Optimizing Robotic Manufacturing in Industry 4.0: A Hybrid Fuzzy Neural Bayesian Belief Networks. Spectrum of Mechanical Engineering and Operational Research 2 (1), pp. 191–203 (2025). https://doi.org/10.31181/smeor21202543