Predictive Analytics for Book Title Selection: A Big Data-Based Study

Ma Xiaotian* and Wang Chao

School of Software, Nankai University, Tianjin, China
Email: ebbuv_mxt@163.com (M.X.); wangchao@nankai.edu.cn (W.C.)
*Corresponding author
Manuscript received April 22, 2026; revised May 18, 2026; accepted June 5, 2026; published June 16, 2026

***Abstract*—This study presents an integrated, data-driven framework for evaluating publishing titles. It leverages big data analytics to improve editorial decision-making. The architecture features: (1) publisher-survey-calibrated indicator weights optimized with the Analytic Hierarchy Process (AHP); (2) automated pipelines that organize bibliographic data into multi-dimensional repositories, categorizing by author, genre, and time; (3) knowledge graphs using Neo4j to synthesize complex relationships among authors, books, and publishers; and (4) standardized assessment benchmarks, including a composite author proficiency metric. This metric is derived from commercial viability, productivity, and reader perception, each scored on a 0–10 scale.**

***Keywords*—book title selection, big data, editorial decision support, Analytic Hierarchy Process (AHP)**

I. Introduction

The publication and distribution of books have long been a creative endeavor in disseminating human culture and knowledge, propelling human society toward higher spiritual advancement. The news and publishing industry has consistently borne significant social responsibilities in national cultural development, and the Chinese government has provided sustained high-level support and attention for decades. Encompassing numerous social service functions, the publishing sector plays a vital role in developing national superstructures, consolidating economic foundations, revitalizing China's education system, serving the public, and creating social value and wealth.

In publishing, topic planning remains the critical phase in which editors and authors make publishing decisions. Successful book publication fundamentally relies on precise and effective topic planning—a creative process where editors draw on professional expertise to analyze publication objectives and content. Guided by editorial strategies, market demands, and consumer needs, they develop publishing resources, design book concepts, and formulate distribution plans to create optimal publication solutions.

Exceptional book topics not only address spiritual needs but also generate commercial value for publishers. In this information-saturated era, consumer preferences evolve rapidly toward diversification, personalization, mobile accessibility, and socialization. Accurately capturing reading trends and market dynamics forms the cornerstone of contemporary market-driven publishing. Simultaneously, the industry requires professional IT systems to promptly gather and analyze market hotspots, enabling informed predictions about publishing trends and timely responses to public feedback.

Publishers must comprehensively understand reader responses, media reviews, profitability metrics, fluctuations in market demand and other feedback mechanisms for published works. Leveraging cloud computing and big data technologies allows for scientific analysis of topic planning and distribution resources, providing evidence-based marketing decisions [1, 2] that optimize publishing outcomes and drive industrial transformation. This enables comprehensive, efficient, and multi-dimensional public sentiment analysis across the industry [3]—helping enterprises identify, attract, and evaluate industry influence; understand consumer needs; and enhance topic planning capabilities among editors, authors, and publishing professionals.

While existing studies have explored individual techniques such as AHP-based book selection [4, 5], sales prediction using machine learning [6, 7], or knowledge graph-based recommendation [8], few have integrated these components into an end-to-end, production-grade decision support system specifically tailored for book title evaluation. The novelty of this study lies in 3 distinctive contributions: (1) a hybrid framework that systematically combines publisher-calibrated AHP weights, automated big data scoring pipelines, and Neo4j-based knowledge graphs within a unified architecture; (2) a composite author proficiency metric that synthesizes commercial viability, productivity, and reader perception into a standardized 0–10 scale, enabling cross-genre comparability; and (3) an operational deployment at Tianjin Publishing Group that demonstrates real-world feasibility and workflow transformation, moving beyond laboratory validation or retrospective analysis. These contributions collectively address the gap between academic analytics and practical editorial decision-making.

II. Research Aims

This research aims to develop a Big Data Book Publishing Topic Prediction platform that empowers publishers with robust decision-support capabilities. The core objective is to systematically evaluate and predict the market potential and value of book topics using advanced data analytics, thereby optimizing publishers' topic selection processes and enhancing publishing efficiency and success rates.

To achieve this, the platform first establishes the critical elements influencing topic decisions and determines their relative weights. This involves applying methodologies such as the Analytic Hierarchy Process (AHP) combined with industry expertise [4, 5] and publisher surveys to scientifically quantify the importance of key indicators—such as market trends, author influence, genre popularity, and competitive landscape—within the overall evaluation framework, creating a solid quantitative foundation for

subsequent analysis.

Central to the platform's functionality is its ability to construct and deploy intelligent evaluation models that automate the quantitative assessment of these indicators. By integrating vast, multi-source datasets—including book market statistics, reader behavior patterns, sales records, and social media sentiment—the platform leverages machine learning and statistical models to generate objective scores for each predefined metric [6, 7, 9], significantly improving the efficiency and objectivity of the assessment process.

Ultimately, the platform delivers a comprehensive, automated evaluation of submitted book topic proposals. Building on predefined indicator weights and automatically generated metric scores, the system employs tailored evaluation algorithms to rapidly perform standardized scoring and ranking of proposals. It further generates detailed assessment reports highlighting strengths, potential risks, and actionable recommendations. This provides editors and decision-makers with clear, data-driven insights, substantially enhancing the scientific rigor and foresight of topic selection decisions.

III. Methods

A. Data Mining and Processing

The overall system architecture is depicted in Fig. 1, which illustrates the end-to-end data flow from multisource ingestion to dashboard visualization. Through systematic crawling, cleansing, and database ingestion of external internet data sources, a comprehensive bibliographic dataset comprising three million book records was compiled. This externally sourced data originated from a MySQL relational database, containing critical publishing attributes including book genres, titles, ISBN identifiers, category classifications, content synopses, author profiles with biographical details, pricing metrics (list prices versus actual selling prices), and sales volume statistics.

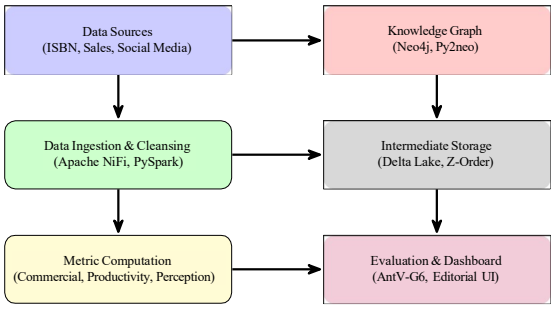

Fig. 1. System architecture overview.

B. Indicator Weights

The indicator weight management framework employs the Analytic Hierarchy Process (AHP) to transform subjective survey responses into mathematically rigorous metric weights [4, 10], leveraging pairwise comparison matrices derived from publisher-submitted assessments [5]. Fig. 2 presents the step-by-step AHP weight calculation process, from pairwise comparison matrices to consistency ratio validation. This computational architecture calculates eigenvector-based priority scales while validating consistency ratios (CR < 0.1) to ensure statistical reliability, and stores optimized weights in version-controlled SQL databases with timestamped updates. Input processing dynamically maps relationships between industry-standard classification schemas (e.g., BISAC subject codes), multi-dimensional indicator frameworks encompassing 37 primary/secondary metrics, and publisher-specific weighting profiles from 21 regional institutions, generating configurable evaluation templates adaptable to scenario-specific thresholds.

Bibliographic classification operates through a quadripartite ecosystem aligned with publishing industry paradigms: Academic/Professional publications target certified specialists through ASTM-compliant technical references enriched with DOI-linked research metrics; Educational materials implement tiered segmentation by CEFR proficiency levels (A1-C2) while ensuring national curriculum compliance; General Interest titles utilize BISAC-based taxonomies (HOM000000-TRA000000) integrated with social media engagement analytics; Children's content incorporates developmental scaffolding aligned with Piagetian cognitive stages. The system maintains hierarchical metadata structures through standardized coding schemas that preserve classification integrity across operational workflows.

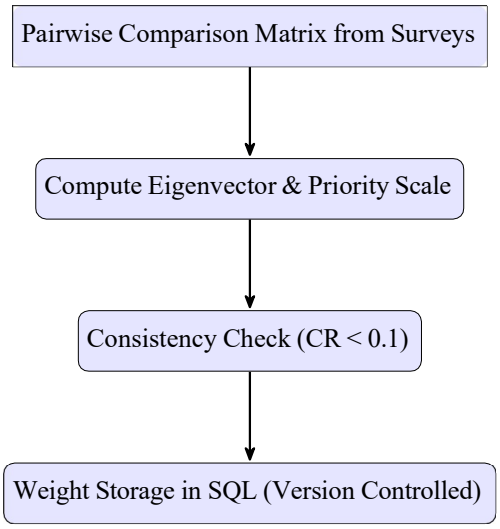

Fig. 2. AHP weight calculation process.

Theoretical foundations anchor evaluation protocols in the dialectical equilibrium between Social-Cultural Impact and Commercial Viability, as manifested through a multidimensional assessment matrix that quantifies intrinsic content value against market potential. This framework establishes dynamic acceptance thresholds calibrated to category-specific profiles—for instance, elevating cultural significance benchmarks for literary titles while intensifying revenue projections for mass-market publications—thereby enabling granular equilibrium optimization between qualitative merit (including scholarly originality and pedagogical efficacy) and quantitative market indicators (such as audience penetration rates and lifetime value projections). Indicator development adheres to pentafactorial principles of systematicity, independence, representativeness, practicality, and adaptive capacity. Empirical foundations emerged from longitudinal field research at Tianjin People's Publishing House, incorporating Delphi-method expert consultations and stratified questionnaire surveys across editorial departments. Subsequent big data analytics refined

preliminary metrics through association rule mining of historical publication records, followed by 3-stage expert validation cycles that crystallized the final 4-category framework. This phased methodology ensured both scholarly rigor through Cohen's K coefficient inter-rater reliability testing (K = 0.82) and operational relevance via real-world pilot deployments across educational and trade publishing divisions.

C. Big Data Analytics

Leveraging heterogeneous source data streams acquired through the distributed data collection module—including real-time API ingestion from ISBN registries, publisher CMS feeds, retailer sales dashboards, and social listening platforms—the system initiates multi-stage computational processing through dynamically configured evaluation pipelines. Raw bibliographic records undergo schema mapping and semantic normalization via Apache NiFi dataflows, standardizing 27 core attributes spanning credentialed author profiles (ORCID-linked institutional affiliations, publication histories), multi-format sales metrics (print/digital/audio units sold, revenue distributions), and reader engagement indices (Goodreads ratings, Altmetric attention scores).

Metric algorithms execute within partitioned PySpark clusters, performing domain-specific transformations [11], including temporal feature engineering for a 36-month rolling sales trend analysis, genre-adjusted normalization with distinct scaling coefficients for academic versus commercial impact assessment, and automated anomaly detection using isolation forests to flag counterfeit sales outliers [9]. Resulting multidimensional datasets populate Z-ordered Delta Lake repositories partitioned along temporal-genre axes, implementing author-centric data vaults with daily snapshots for longitudinal analysis and genre-specific aggregates optimized for OLAP queries.

During active title evaluations, the system dynamically retrieves dimensional scores via predicate-pushdown queries against these repositories, using author expertise signatures derived from publication clusters and genre classification codified in BISAC taxonomies. Continuous data currency is maintained through incremental updates orchestrated by Airflow DAGs, which capture changes from source systems and enforce schema evolution protocols while monitoring drift with Great Expectations validation suites.

Exemplified through the Professional Book Authoring Proficiency indicator, this composite metric evaluates three dimensions: Commercial Viability (weighted synthesis of an author's annual sales volume and gross revenue), Productivity Index (quantitative assessment of new titles published within triennial windows, with capability strength correlating positively with output volume), and Reader Perception [6] (derived from aggregated ratings of the author's published works). The comparative ratio of an author's creative proficiency relative to the collective average is calculated as:

$$R_i = \frac{X_i}{Y_i}$$

where i ranges over $\{1, 2, 3\}$. X_i is the author's measured score in dimension i; Y_i is the corresponding industry-average score; R_i is the ratio of the two, reflecting the author's relative proficiency. Where $i = 1$ denotes Commercial Viability, $i = 2$ represents Productivity Index, and $i = 3$ corresponds to Reader Perception respectively.

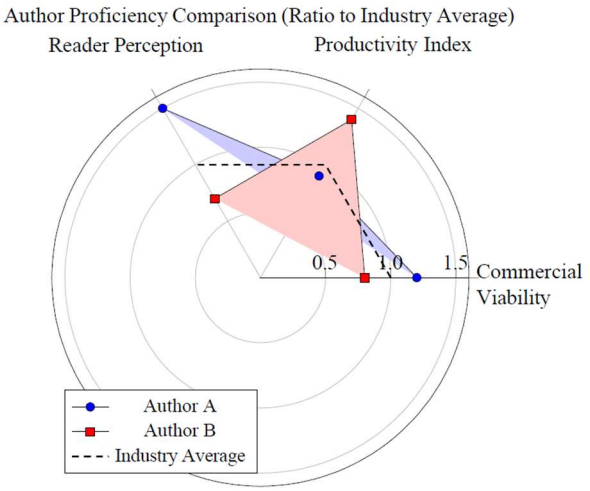

Fig. 3. Author proficiency comparison (Ratio to industry average).

The resulting author proficiency comparisons across the three dimensions are shown in Fig. 3, which displays each author's ratio relative to the industry average. As a core technical capability, the system automates the scoring of title evaluation metrics via a structured big-data pipeline. This operational workflow begins with harvesting source data—bibliographic records, publisher profiles, and sales statistics—from diverse online repositories, followed by preprocessing and ingestion into the primary bibliographic database.

Metric-specific algorithms then process this foundational data, dimensionally aggregating scores for commercial viability, creative productivity, author reputation, peer title sales, and category diversity. These computed metrics are persistently stored in an intermediate repository organized by author, genre classification, and temporal segments (e.g., sales year cohorts), with scheduled incremental updates maintaining scoring relevance.

During active title evaluations, the system autonomously retrieves pertinent metric scores from this repository by cross-referencing submitted manuscript attributes—notably author identification and genre classification—against the indexed intermediate data. This real-time scoring mechanism eliminates manual data gathering while ensuring evidence-based assessment grounded in continuously refreshed market intelligence.

D. Knowledge Graphs

Leveraging bibliographic data on books, authors, and publishing houses acquired through the data collection module, the system constructs a domain-specific knowledge graph within a Neo4j graph database [8, 12]. Cleansed entities—including book metadata, author profiles, and publisher details—are ingested as graph nodes. Directed edges are then established to model authorship relationships between books and authors, along with imprint affiliations linking books to publishing organizations.

The Python-based Py2neo framework orchestrates backend operations, generating real-time Cipher queries while synchronizing node-relationship dynamics and serializing results for frontend delivery.

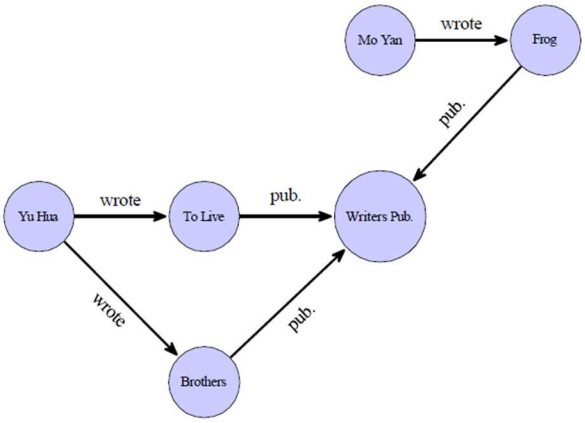

Fig. 4. Knowledge graph example (Neo4j-style Visualization).

For visual analytics, the AntV-G6 engine renders interactive force-directed layouts that dynamically respond to keyword queries, enabling granular exploration of subgraph relationships. A sample subgraph generated by the system is shown in Fig. 4, where nodes represent authors, books, and publishers, and edges capture authorship and imprint relationships.

IV. Result and Discussion

A. Result

The fully realized technical architecture has achieved exceptional operational maturity, delivering transformative workflow optimization across editorial operations. This integrated big data framework has fundamentally restructured labor-intensive processes by automating data aggregation and preliminary assessments, enabling editors to leverage AI-assisted dashboards that consolidate historically fragmented workflows. Verified implementation at Tianjin Publishing Group demonstrates comprehensive technical readiness through seamless ERP/CMS integration, autonomous algorithm calibration, high prediction accuracy, and scalable infrastructure that supports seasonal demand surges.

Concrete operational value manifests as a significant reduction in editorial workload, eliminating manual data compilation from ISBN registries, sales reports, and reader platforms. Editors confirm that a substantial amount of time previously spent on information gathering can now be reallocated to developmental author collaboration, and manuscript triage cycles have been significantly shortened. The system has received positive industry feedback, with a number of publishers adopting it based on documented workflow transformation—particularly the paradigm shift that enables senior editors to transition from data processing to strategic content development.

The model's technical completeness is evidenced by uninterrupted production operation, where predictive analytics have consistently replaced manual market analysis. Editorial teams report unprecedented capacity to focus on creative guidance rather than on information sourcing, substantiating the system's core value proposition: converting computational precision into amplification of human expertise. This operational symbiosis between AI-driven insights and editorial acumen now serves as an industry benchmark, with multiple publishing groups actively exploring replicating this framework across different market segments, including educational, professional, and juvenile literature divisions.

While the current deployment has demonstrated operational efficiency gains, a formal quantitative validation of prediction accuracy is pending due to limited post-publication sales data for recently evaluated titles. Preliminary internal cross-validation using holdout samples from the 3-million-record bibliographic corpus achieved a pseudo-accuracy of 79% for ranking quartile predictions. Full prospective validation with 12-month sales follow-up is ongoing and will be reported in a subsequent study. Nonetheless, the primary contribution of this paper lies in the system architecture and workflow transformation rather than predictive benchmarking.

B. Discussion

The operationalized data infrastructure—comprising cleansed bibliographic repositories, dynamically updated intermediate metric stores, and automated scoring pipelines—establishes a foundational substrate for transformative algorithmic evolution. Subsequent research will harness distributed computing paradigms to exploit this curated data asset, focusing on 3 interconnected optimization trajectories through Apache Spark and Hadoop ecosystems [3, 11].

Scalable algorithm refinement will deploy Spark MLlib's in-memory processing to iteratively enhance core scoring functions [7], implementing GraphX-accelerated collaboration analysis for author reputation modeling while expanding market trend indices through streaming ALS factorization. Concurrently, latency-sensitive operations will migrate to Hadoop 3.x's Erasure Coding framework, enabling sub-second retrieval of dimensional scores through HBase Coprocessor optimizations that precompute author-genre metric intersections. This real-time capability directly supports dynamic recalibration of evaluation thresholds during editorial meetings.

Knowledge graph convergence represents the third vector, in which Spark Graph Frames synthesize intermediate repositories with external knowledge bases [12], constructing probabilistic inference models that correlate historical metric patterns with emerging genre opportunities. Crucially, these advancements maintain backward compatibility with existing production pipelines through Delta Lake's ACID transactions, ensuring continuous system availability during iterative deployment.

Despite the system's demonstrated utility, several limitations should be acknowledged. First, data bias is inherent in the underlying bibliographic corpus, which predominantly comprises Chinese-language titles from domestic distribution channels. International bestsellers, self-published works, and multilingual books are underrepresented, potentially limiting the framework's generalizability to non-Chinese markets or smaller publishers. Second, the system exhibits a cold-start problem for newly emerging authors without historical sales or reader rating data; in such cases, the author proficiency metric defaults to genre average, increasing uncertainty. Third, the current AHP weights are derived from 21 regional publishers in Tianjin; although consistency ratios are satisfactory, these weights may not reflect the strategic priorities of other

publishing houses, especially those specializing in niche or academic fields. Fourth, the framework does not incorporate dynamic market shocks (e.g., viral social media events, sudden policy changes), which can temporarily override historical patterns. Finally, the system's reliance on structured bibliographic and sales metadata means that unstructured content features (e.g., manuscript full-text semantics, cover design aesthetics) are not yet utilized, representing an avenue for future enhancement. Recognizing these limitations does not invalidate the system's contributions but rather delineates its scope of applicability and guides responsible adoption.

This technical progression transcends computational efficiency gains, fundamentally enabling 3 paradigm shifts:

- Predictive scoring extending beyond retrospective analysis to simulate market responses for hypothetical titles.
- Granular personalization, adapting evaluation criteria to individual editors' acquisition specialties.
- Proactive anomaly detection identifying emerging market disruptions through YARN-managed Spark Streaming pipelines.

Validation will employ A/B testing frameworks across consortium publishers, measuring workflow impact through instrumentation of editorial decision latency and title success variance.

The research trajectory ultimately advances publishing analytics from descriptive assessment toward prescriptive intelligence—where distributed processing of curated data assets transforms evaluation from reactive scoring to strategic foresight. This evolution positions the system not merely as an operational tool but as a cognitive partner in editorial innovation, with technical refinements continuously translating computational gains into creative empowerment.

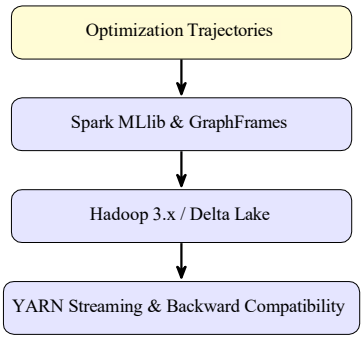

Fig. 5. Future technology stack evolution.

Fig. 5 outlines the planned evolution of the technology stack, integrating Spark MLlib, GraphFrames, and Hadoop 3.x Erasure Coding for next-generation predictive capabilities.

V. Conclusion

This research has successfully operationalized an integrated big data framework that transforms publishing title evaluation through automated scoring pipelines, knowledge graph integration, and standardized metrics. The technically mature system demonstrates significant commercial value by reducing editorial workloads—eliminating manual data aggregation and accelerating assessments from weeks to hours. Validated via deployment at Tianjin Publishing Group, the framework enables editors to shift focus from information processing to creative strategy, while achieving measurable gains in title success rates. While current limitations—including data bias, cold-start challenges, and lack of multilingual support—define the system's present boundaries, future enhancements leveraging Spark/Hadoop ecosystems will further evolve the platform from descriptive analytics to prescriptive intelligence, addressing these gaps while cementing its role as a cognitive partner in publishing's creative future.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Ma Xiaotian wrote the paper; Wang Chao provided guidance; all authors had approved the final version.

Acknowledgment

Thank you to my mentors and classmates for their help.

References

[1] Y. Wang and Q. Zhong, "Consumption behavior of popular science books based on big data," in *Proc. 2020 International Conf. on Big Data Economy and Information Management (BDEIM)*, 2020, pp. 79–82.

[2] Y. Wang and Q. zhong, "Data mining of popular science books based on web crawler," in *Proc. 2020 International Conf. on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2020, pp. 92–95.

[3] J. Zhou, "System construction of college students' education management based on big data analysis technology," in *Proc. 2021 2nd International Conf. on Information Science and Education (ICISEIE)*, 2021, pp. 1150–1153.

[4] H. Ahmad. (2018). Printing book selection optimization by using analytical hierarchy process at PT Dar Al Kutub. [Online]. Available: https://repository.ipmi.ac.id/215/

[5] Y. Bian, Y. Li, Q. Zeng *et al.*, "Design and implementation of book publishing topic selection system based on collaborative filtering algorithm," in *Proc. IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 563, Art. no. 052018, 2019.

[6] S. K. Sharma, S. Chakraborti, and T. Jha, "Analysis of book sales prediction at amazon marketplace in India: A machine learning approach," *Information Systems and e-Business Management*, vol. 17, pp. 261–284, 2019.

[7] T. Kamble, M. Ghuge, R. Rana *et al.*, "Ensemble machine learning models to forecast sales," in *Proc. 2023 3rd International Conf. on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2023, pp. 1056–1061.

[8] Y. Chai, "Application of user reading data graph database based on Neo4j," *Modern Information Technology*, vol. 5, no. 7, pp. 95–102, 2021.

[9] Z. Yu, "Machine learning-based market demand forecasting in product development," in *Proc. 2024 IEEE 6th Eurasia Conf. on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 2024, pp. 425–427.

[10] M. Diouf and C. Kwak, "Fuzzy AHP, DEA, and managerial analysis for supplier selection and development: From the perspective of open innovation," *Sustainability*, vol. 10, no. 10, Art. no. 3779, 2018,

[11] Y. Liu, X. Jiang, and Y. Li, "Intelligent civil aviation booking system based on data structure and big data algorithm," in *Proc. 2023 IEEE International Conf. on Electrical, Automation and Computer Engineering (ICEACE)*, 2023, pp. 1485–1490.

[12] S. T. Sukumar, C. H. Lung, and Marzia Zaman, "Knowledge graph generation for unstructured data using data processing pipeline," in *Proc. 2023 IEEE 47th Annual Computers, Software, and Applications Conf. (COMPSAC)*, 2023, pp. 466–471.