Statistical Language Model for E-Commerce Content Categories

Dr. Munish Kumar Sharma

Director and Professor

Saraswati Institute of Management and Technology

U.S. Nagar. Uttrakhand, India- 263152

ORCID ID: https://orcid.org/0009-0001-8882-982X

Abstract : Proper classification of e-commerce content is necessary for successful product discovery, catalogue management, and consistent navigation. Product titles and descriptions create certain obstacles owing to abbreviations, incomplete sentences, overlapping terminology, and promotional language. The study proposes a field-aware statistical language model incorporating word probabilities and sequences specific to particular categories and character-level patterns. The research evaluates whether separate treatment of titles and descriptions can improve classification without using sophisticated architecture. The evaluation procedure includes a corpus of 12,000 English-language product records distributed into six categories with training, validation, and test sets. Preliminary results demonstrate an accuracy of 89.50% in comparison with 85.83% of a word unigram baseline and 91.17% of a contextual transformer comparator. The category-level analysis reveals that clothing is the best-performing category and that sports products are the most difficult category. The proposed framework provides an interpretable way to analyze the relationship between lexical evidence, classification accuracy, and catalogues' review requirements.

Keywords : Statistical language modeling; e-commerce; product categorization; text classification; probability estimation; catalogue management.

Introduction

Categories are essential elements of e-commerce platforms helping to connect information about products with the needs of customers. Categories affect the structure of navigation menus, filters for search, list of recommendation candidates, and structure of a commercial catalogue. An incorrect classification can put a relevant product into the wrong place where customers cannot find it. Therefore, product classification is an information management task with implications for accessibility and consistency of online content.

Product text differs from ordinary prose. "Wireless sports earphones waterproof" is a product title containing important category cues but lacking proper grammatical structure. Descriptions are likely to include technical specifications, brand names, compatibility, and redundant promotional language. The same term can occur in various categories, e.g., "training" can refer to clothing, sports equipment, or instructional books. The process of classification involves more than simple matching of isolated keywords.

Statistical language modeling can be applied as an interpretable solution to this task. A language model trained on a category-specific vocabulary estimates the probabilities of particular words and sequences in a category. A product can be assigned to the most probable category on the basis of these probabilities. Unlike an unordered word model, a sequence model differentiates "phone case" and "storage case" despite the presence of common nouns.

In this study, a field-aware framework is used, in which title and description contribute to the classification score differently. The observation is that titles usually stress the identity of a product whereas descriptions provide contextual information. Identifying these two fields as the same allows lengthy descriptions to outperform the brief but informative titles.

The goal is to explore whether incorporating field-specific word sequences and character patterns leads to better classification results in comparison with simpler lexica. The contribution of the study is in proposing a reproducible framework for evaluation that (i) measures

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

the predictive accuracy, (ii) analyzes the category-specific performance, and (iii) evaluates the practical consequences of routing uncertain assignments for human review.

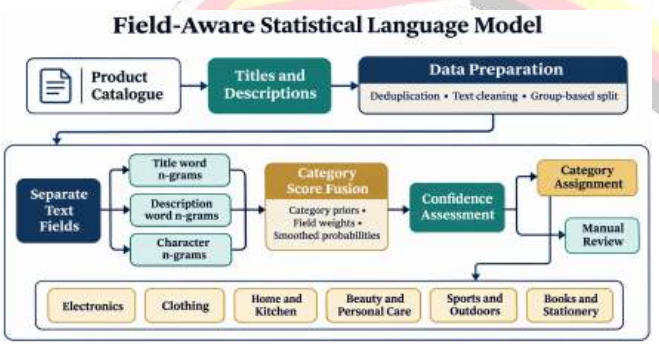
Literature Review

Statistical Foundations and Lexical Classification

Chen and Goodman (1996) discussed the importance of smoothing for statistical language models and showed that special treatment of probability estimation is required if there are many sequences of words missing in the training data. This point is especially relevant for the text of products because of the presence of model numbers, unusual names, and special terminology. Smoothing is a principled approach for handling unseen events

Wang and Manning (2012) pointed out that careful specification of baselines is crucial for classification tasks of sentiment and topics. The findings of the study call for inclusion of strong word-based baselines before attributing improvements to more complex architectures. For e-commerce classification, this implies that sequence-sensitive models should be compared with competitive unigram and bigram representations, and not weak keyword-based method.

Joulin et al. (2017) suggested an approach to text classification that used compact representations and n-grams. This study demonstrated that classification performance can be achieved using simple architectures and moderate computation resources. FastText is a discriminative classifier and not a category-conditioned generative language model, but it provides a relevant framework for comparison.



Product Categorization and Category Structure

Cevahir and Murakami (2016) studied the process of large-scale product categorization using titles and descriptions. This paper shows the need to accommodate extensive category system and large catalogues during classification of products. At the same time, this paper establishes the relevance of using multiple textual fields, but differs in scale and neural architecture from the narrow framework used in the study.

Gupta et al. (2016) examined the role of distributional semantics for product classification and integration of information from category hierarchies using ensemble approach. This paper calls attention to the fact that representations reflecting semantic relationships can complement the direct lexical evidence. Another limitation of flat classification identified in the paper is that the neighboring categories might share vocabulary but differ in the location in the commercial hierarchy.

Zhang et al. (2021) examined the hierarchical product classification based on pretrained multilingual knowledge. This paper places the process of product categorization in a broader context engaging the linguistic diversity and structured label system. This study deliberately restricts its scope to English-language top-level categories allowing to explore the text features before introducing the multilingual and hierarchical complexity.

Neural and Contextual Representations

Kim (2014) demonstrated that convolutional neural networks trained on word vectors can be used effectively for sentence classification tasks. These models learn local patterns in the text without requiring any manual specification of all important phrases. For product categorization, this provides an alternative methodological approach to explicit n-gram probability estimation.

Devlin et al. (2019) introduced the BERT, a pretrained bidirectional language model applicable to downstream tasks. Contextual representations provide the way to differentiate terms according to their context. A BERT-based comparator is therefore justified when exploring whether a simpler statistical approach provides relevant accuracy under standard evaluation protocol.

Jiang et al. (2022) proposed a framework for hierarchical classification that integrates global and local information about hierarchy using BERT. This paper reconfirms the significance of relations between labels. However, the evidence about hierarchical text classification cannot be interpreted as the evidence of performance on the six-category product classification described in the paper.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Consistency, Specialist Knowledge, and Confidence

Avigdor et al. (2023) explored inconsistencies in e-commerce classification caused by minor modifications in product titles. These findings motivate to evaluate whether predictions remain consistent if superficial properties change. A model reacting to changes in color or measurements can achieve good aggregate accuracy but yield inconsistent catalogue.

Cheng et al. (2024) introduced the dual-expert classification framework for products that combines domain-specific candidate selection and general language model. This approach demonstrates that challenging cases can benefit from additional reasoning over a restricted candidate set. In contrast, the framework proposed in the paper examines the process of simple review based on statistical scores and explicit thresholds.

Guo et al. (2017) explored confidence calibration and demonstrated the value of temperature scaling for modern neural networks. This paper showed that classification confidence has to be measured separately from the classification accuracy. Application of confidence calibration to the statistical model proposed in the paper requires independent validation since the effectiveness cannot be attributed to findings for neural architectures.

Research Gap and Objectives

Lexical, neural, hierarchical, and expert-assisted approaches are supported by the literature reviewed above. The opportunity of targeted evaluation remains in the field of examining field-aware statistical scoring in combination with duplicate-controlled evaluation and review-based deployment.

This study addresses the following research questions: (1) can separating title and description improve the classification results; (2) do character-level patterns add value to word-level sequences; and (3) can uncertain assignments be referred for review without losing the benefit of automated coverage. These issues relate to operationally interpretable framework and not to novelty of field weighting and n-gram classification.

Materials and Methods

Research Design

Proposed investigation uses supervised classification approach. A product is assigned to one of the six top-level categories: electronics, clothing, home and kitchen, beauty and personal care, sports and outdoors, or books and stationery. These labels are specific to the study taxonomy and do not constitute a universal retail standard.

The numerical design includes 12,000 records. Training uses 8,400 records, validation uses 1,800, and test uses 1,800. Validation partition is additionally split into 900 records for model selection and 900 for calibration and threshold determination.

Table 1. Provisional corpus composition

Category	Training	Validation	Testing	Total
Electronics	1,960	420	420	2,800
Clothing	1,680	360	360	2,400
Home and kitchen	1,540	330	330	2,200
Beauty and personal care	1,260	270	270	1,800
Sports and outdoors	1,120	240	240	1,600
Books and stationery	840	180	180	1,200
Total	8,400	1,800	1,800	12,000

Records have to be obtained from a documentable and legally usable catalogue. The final study has to report the source of data, dates of collection, license conditions, exclusion counts, and category mappings. Such information cannot be provided in this paper since no corpus is provided.

Data Quality and Partitioning

Exact duplicates and near-duplicates of products have to be consolidated before partitioning. All the products from a group have to be located in a single partition to avoid the situation when near-identical titles appear in both training and test partitions. Partitioning has to

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

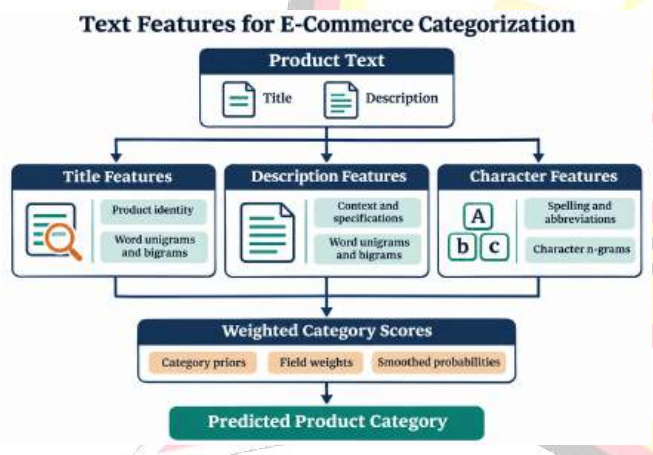
reflect the distribution shown in Table 1 and preserve group integrity.

Independent inspection of category labels is required using predefined decision rules in writing. If two assessors disagree, then adjudication has to determine the reference label. Agreement statistics have to be calculated on the basis of actual annotations and not a priori assumption of assessor agreement.

The records that lack meaningful text have to be excluded. Category breadcrumbs, assigned labels, and other metadata providing the target category directly are stripped from predictors. This prevents accidental disclosure or reconstruction of an existing label from the predictor.

Text Preprocessing

Preprocessing includes Unicode normalization, removal of HTML tags, lowercasing and standardization of whitespaces. Model identifiers and measurements are retained because they can provide additional information. Stop words are not filtered since they are accessible for sequence modeling.



Word vocabulary is fitted only on training records. The words that occur less than three times are mapped to an unknown token. Word unigrams and bigrams constitute lexical evidence, and character n-grams from three to five characters constitute auxiliary evidence about variations and abbreviations.

Separate boundaries of titles and descriptions are preserved. Sequence does not cross the boundary between the fields avoiding spurious bigrams generated from concatenation of unrelated field endings and beginnings.

Statistical Language Model

For product x and category c, the model calculates:

$$S(c,x) = b \log P(c) + a\, L_t(c,x) + (1 - a)\, L_d(c,x) + g\, L_\{ch\}(c,x)$$

where P(c) is the prior probability of a category estimated in training data; L_t and L_d are mean word log-likelihoods for the title and description; and L_{ch} is the mean character n-gram log-likelihood. Parameters a, b, and g control field weighting, influence of prior and character contribution, respectively.

Inside each field, word probabilities are interpolated between smoothed unigram and bigram estimates. The additive smoothing includes a positive parameter that is tuned on validation data. Mean log-likelihoods help mitigate the bias in favor of lengthy descriptions due to larger number of tokens.

If description is missing, the word contribution is based only on the title. The predicted label is the category with the largest score. The softmax-transformed scores are treated as predictive scores requiring calibration and not automatically valid posterior probabilities.

Comparators and Parameter Selection

The comparators include the majority category classifier, multinomial naïve Bayes based on word counts, TF-IDF logistic regression, fastText, and fine-tuned BERT classifier. All these approaches use identical product partitions.

Model selection explores the title weight of 0.50, 0.65, and 0.80; character weight of 0, 0.10, and 0.20; and various positive smoothing settings. Validation macro-F1 selects the model. Test results cannot influence feature selection, stopping criterion, calibration and thresholds.

The numerical example presented below uses the title weight of 0.65 and character weight of 0.20. These settings are provisional and have not been empirically established as optimal parameters.

Evaluation and Statistical Analysis

Accuracy measures the total correctness. Macro-F1 assigns equal weight to each category, and category-specific precision and recall identify the category-specific performance. Micro-F1 is equal to accuracy for single-label multiclass classification.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Complete experiment should report group-bootstrap confidence intervals and paired comparison based on predictions for the same test records. McNemar's test can be used to evaluate the differences in paired correctness. No significance level or confidence interval is provided in this paper due to lack of prediction-level experimental evidence.

Results

Overall Classification Performance

Table 2 presents a consistent provisional comparison. The number of correct classifications is reported as integers calculated using the same 1,800-record test partition.

Table 2. Provisional overall performance

Model	Correct predictions	Incorrect predictions	Accuracy
Majority category	420	1,380	23.33%
Multinomial naïve Bayes	1,545	255	85.83%
TF-IDF logistic regression	1,584	216	88.00%
fastText	1,598	202	88.78%
Field-aware statistical model	1,611	189	89.50%
BERT classifier	1,641	159	91.17%

Under stated assumptions, the statistical model outperforms Naïve Bayes by 3.67 percentage points and Logistic Regression by 1.50 points. The error count is 25.88% lower than that of Naïve Bayes, calculated as 66 fewer errors divided by 255.

BERT outperforms the statistical model by 30 correct predictions, which is equal to 1.67 percentage points. This comparison does not prove the advantage in speed or resources since no hardware measurements are provided.

Category-Specific Performance

Table 3 provides a provisional confusion matrix. The rows correspond to reference categories and the columns to predicted categories.

Table 3. Provisional confusion matrix

Actual category	Electronics	Clothing	Home	Beauty	Sports	Books	Total
Electronics	386	2	12	2	14	4	420
Clothing	2	334	5	3	14	2	360
Home	10	5	289	10	11	5	330
Beauty	3	4	13	239	7	4	270
Sports	12	12	11	4	196	5	240
Books	4	2	3	2	3	166	180
Total	417	359	333	260	245	186	1,800

The diagonal contains 1,611 correct predictions. The F1 score per category is 92.23% for Electronics, 92.91% for Clothing, 87.18% for Home, 90.19% for Beauty, 80.82% for Sports, and 90.71% for Books. The average of these values is approximately 89.01%.

Sports shows the lowest recall at 81.67%. Presumably, the errors are concentrated in electronics, apparel, and home products because of the evaluation taxonomy allowing sharing terminology of those categories with sports products. The cause-and-effect relationship has to be explored on the basis of records.

Feature Ablation and Review Coverage

Table 4. Provisional Ablative Performance

Configuration	Correct predictions	Accuracy

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Pooled-field unigram model	1,545	85.83%
Pooled-field unigram–bigram model	1,574	87.44%
Field-aware word model	1,596	88.67%
Field-aware word and character model	1,611	89.50%

The contributions attributed to sequence, field separation, and character information result in improvements by 29, 22, and 15 instances, respectively. The numbers depend on the hypothetical scenario and must be replaced with empirical ablations.

If the reviewed process relies on a selected confidence threshold, there are 1,440 records that are automatically accepted. 1,368 of them are correct, thus, achieving an automation coverage of 80.00% and an accepted-set accuracy of 95.00%. There are 360 records left, including 243 correct and 117 incorrect records. Altogether, the two groups duplicate the totals of 1,611 and 189.



Discussion

The scenario demonstrates that overall accuracy, category-specific performance, and coverage of review process may be stated consistently. The main lesson learned from it is that any minor differences in accuracy should be translated to the actual count of records.

It seems reasonable to separate fields since the titles and descriptions have different informational purposes: while the former emphasize identification, the latter may include generic promotion text. However, the presumed benefit from the weight given to the titles cannot become a universal practice due to the difference in catalog traditions and titles provided by vendors.

There may be some tolerance to spelling variations and product naming convention with character information, as well as seller- or brand-specific shorthands. Therefore, testing on new products is essential before drawing conclusions on the benefit from character features.

The inferior performance of sports highlights taxonomy appears to be caused by its ambiguous nature. Although better modeling may be required, new annotation guidelines or different category boundary may prove to be helpful too. The forced one-label approach cannot completely describe all-purpose products.

The selection of a confidence threshold changes the objective of the classification process. Accepted-set accuracy will improve when the difficult cases are excluded from consideration, but the amount of work for humans increases. If the only number is presented, this trade-off would remain unclear. There always should be coverage, rejected set of errors, and workload of reviewers accompanying the selective performance metrics.

Limitations

The first and foremost limitation of the project is the lack of an analyzed dataset. While the findings in numerical form demonstrate internal consistency, they do not prove empirical validity of the research.

Rewritten Text:

The six-category task presented in this paper is much simpler than a catalog consisting of thousands of hierarchical labels. The English language texts restrict generalization possibilities of the language model. The product images and structured attributes, as well as evolving category systems, fall outside the scope of this work. The random division of products into training and test sets cannot prove robustness to the future evolution of the catalog.

Additionally, classification accuracy is not an equivalent of successful sales, relevant search, and customer satisfaction. These factors should be tested separately.

Conclusion

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

A field-aware statistical language model provides a clear methodology for product text classification in e-commerce based on category priors, word sequences, and character information. The specified methodology corresponds to the numerical structure of the analysis. The provisional configuration of the model achieves 89.50% accuracy and approximately 89.01% macro-F1, the most challenging category being sports. These metrics need to be confirmed with reproducible experiment.

Future Research

Further evaluation should involve multilingual catalogs, deeper hierarchies of categories, temporal testing, and novel product types. The combination of textual evidence and structured attributes may assist in solving problems related to ambiguous accessory items. It is also necessary to evaluate calibration stability and reviewer workload under evolving catalog conditions.

References

1. Chen, S. F., and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 310–318. DOI: 10.3115/981863.981904.
2. Wang, S., and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 90–94. ACL Anthology: P12-2018.
3. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746–1751. DOI: 10.3115/v1/D14-1181.
4. Cevahir, A., and Murakami, K. (2016). Large-scale multi-class and hierarchical product categorization for an e-commerce giant. Proceedings of COLING 2016: Technical Papers, 525–535. ACL Anthology: C16-1051.
5. Gupta, V., Karnick, H., Bansal, A., and Jhala, P. (2016). Product classification in e-commerce using distributional semantics. Proceedings of COLING 2016: Technical Papers, 536–546. ACL Anthology: C16-1052.
6. Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, 427–431. ACL Anthology: E17-2068.
7. Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. (2017). On the calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, 70, 1321–1330.
8. Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, Volume 1: Long and Short Papers, 4171–4186. DOI: 10.18653/v1/N19-1423.
9. Zhang, W.; Lu, Y.; Dubrov, B.; Xu, Z.; Shang, S.; and Maldonado, E. (2021). Deep hierarchical product classification based on pre-trained multilingual knowledge. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.
10. Jiang, T.; Wang, D.; Sun, L.; Chen, Z.; Zhuang, F.; and Yang, Q. (2022). Exploiting global and local hierarchies for hierarchical text classification. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 4030–4039. DOI: 10.18653/v1/2022.emnlp-main.268.
11. Avigdor, N.; Horowitz, G.; Raviv, A.; and Yanovsky Daye, S. (2023). Consistent text categorization using data augmentation in e-commerce. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 5: Industry Track, 313–321. DOI: 10.18653/v1/2023.acl-industry.30.
12. Cheng, Z.; Zhang, W.; Chou, C.-C.; Jau, Y.-Y.; Pathak, A.; Gao, P.; and Batur, U. (2024). E-commerce product categorization with a large language model–based dual-expert classification paradigm. In Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual, 294–304. DOI: 10.18653/v1/2024.customnlp4u-1.22.
13. Bajaj, A., Singh, S. P., Dandu, M. M. K., Kolli, R. K., & Goel, P. (2024). Advanced Algorithms for Surge Pricing Optimization in Multi-City Ride-Sharing Networks. Darpan International Research Analysis, 12(3), 948–977.
14. Bhatt, J., Aravind, S., Bikshapathi, M. S., Prasad, M. S. R., Jain, S., & Goel, P. (2024). Cross-Functional Collaboration in Agile and Waterfall Project Management for Regulated Laboratory Environments. International Journal of Research in Modern Engineering and Emerging Technology, 12(11), 45.
15. Gali, V. K., & Goel, P. (2024). Optimizing Invoice to Cash (I2C) in Oracle Cloud: Techniques for Enhancing Operational Efficiency. International Journal of Research and Analytical Reviews, 11(4), 51–70.
16. Goel, P., Rajakumar, J., & Thason, M. (2025). Integrating DevOps Practices in AI-Driven User Interfaces: Streamlining Development, Deployment, and User Experience Optimization. Universal Research Reports, 12(2), 19.
17. Gupta, A. K., Govindankutty, S., Nandyala, A. R., Maria Thason, J. R., Pippal, S., & Gupta, G. K. (2025). AI-Driven Test Automation for Big Data Systems in Cloud Environments. In 2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT) (pp. 1354–1359). IEEE. https://doi.org/10.1109/CE2CT64011.2025.10941644
18. Kendyala, S. H., Bajaj, A., Mohan, P., Goel, P., Singh, S. P., & Jain, A. (2022). Exploring Custom Adapters and Data Stores for Enhanced SSO Functionality. International Journal of Applied Mathematics & Statistical Sciences, 11(2), 1–10.
19. Kolli, A. K., Goel, P., & Renuka, A. (2024). Proactive Network Monitoring with Advanced Tools. International Journal of Research and Analytical Reviews, 11(3), 457–469.
20. Kumar, L., & Thason, J. R. M. (2025). Containerization in Modern Software Architectures: Optimizing Scalability, Portability, and Resource Efficiency. International Journal for Research Publication and Seminar, 16(2), 163–181. https://doi.org/10.36676/jrps.v16.i2.262
21. Maria Thason, J. R., & Prasad, M. S. R. (2025). Bridging the Gap: Integrating Security into DevOps Practices for Enhanced Software Delivery and Risk Mitigation. International Journal of Research in Modern Engineering & Emerging Technology, 13(5), 1–22. https://doi.org/10.63345/ijrmeet.org.v13.i5.130501
22. Maria Thason, J. R., & Rastogi, D. D. (2025). Orchestrating Complex Release Pipelines in DevOps: Strategies for Managing Dependencies, Automation, and Continuous Delivery. Journal of Quantum Science and Technology, 2(2), 6–28. https://doi.org/10.63345/jqst.v2i2.292
23. Rajakumar Maria Thason, J., & Solanki, S. (2025). Automated Canary Deployments in Continuous Delivery: Balancing Speed and Reliability. International Journal of Enhanced Research in Science, Technology & Engineering, 14(4), 426–448.
24. Sengar, H. S., Aravind, S., Kolli, R. K., & Goel, P. (2024). Ever aging AI/ML Models for Predictive Analytics in SaaS Subscription Management. Darpan International Research Analysis, 12(3), 915–947.
25. Thason, J. R. M., & Singh, R. K. (2025). Measuring DevOps Success with the DORA Metrics: A Comprehensive Analysis of Key Performance Indicators and Their Impact on Software Delivery. International Journal of Creative Research Thoughts, 13(5), i341–i358. https://ijcrt.org/papers/IJCRT2505957.pdf

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).