International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group | Volume 7 Issue 1 PP 78-85, 2026
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V7I1P112

Original Article

Impact of Text Preprocessing Techniques on SMS Spam Classification Accuracy

Samon Daniel
Ladoke Akintola University of Technology.

Received On: 15/12/2025 Revised On: 16/01/2026 Accepted On: 26/01/2026 Published On: 02/02/2026

Abstract - *Text preprocessing plays a critical role in enhancing the performance of SMS spam classification systems by transforming raw text into a structured and machine-readable format. This study examines the impact of various text preprocessing techniques on the accuracy of SMS spam classification models. Key preprocessing steps analyzed include text normalization, tokenization, stop-word removal, stemming, lemmatization, handling of special characters, and feature scaling. Using benchmark SMS spam datasets, multiple machine learning classifiers are evaluated under different preprocessing configurations to assess their influence on classification accuracy, precision, recall, and F1-score. The results demonstrate that appropriate preprocessing significantly improves model performance by reducing noise, dimensionality, and data sparsity. However, the study also highlights that excessive or improper preprocessing can lead to information loss and reduced accuracy. The findings provide practical insights into selecting optimal preprocessing pipelines for efficient and accurate SMS spam detection systems, particularly in resource-constrained and real-time environments.*

Keywords - *SMS Spam Classification, Text Preprocessing, Natural Language Processing, Machine Learning, Feature Engineering, Classification Accuracy, Spam Detection.*

1. Introduction

1.1. Background of SMS Spam Classification

The rapid growth of mobile communication has led to a significant increase in unsolicited and malicious SMS messages, commonly referred to as SMS spam. These messages are often used for advertising, phishing, fraud, and the distribution of harmful links, posing security and privacy risks to users. SMS spam classification aims to automatically distinguish spam messages from legitimate (ham) messages using computational techniques, primarily leveraging machine learning and natural language processing (NLP). Effective classification systems are essential for improving user experience, reducing exposure to threats, and supporting telecom operators in managing network abuse.

1.2. Importance of Text Preprocessing in NLP-Based Classification

SMS messages are typically short, informal, and noisy, containing abbreviations, misspellings, emojis, special characters, and inconsistent grammar. Text preprocessing is a crucial step in NLP-based classification, as it transforms raw text into a cleaner and more structured representation suitable for feature extraction and model learning. Techniques such as normalization, tokenization, stop-word removal, stemming, and lemmatization help reduce noise, control vocabulary size, and address data sparsity. The quality of preprocessing directly affects feature representation and, consequently, the performance of SMS spam classification models.

1.3. Problem Statement and Research Motivation

Despite the widespread use of text preprocessing in SMS spam detection, there is no universal agreement on which preprocessing techniques or combinations yield optimal classification accuracy. In some cases, aggressive preprocessing may remove discriminative information, while insufficient preprocessing may leave noise that degrades model performance. This lack of clarity motivates a systematic investigation into how different text preprocessing techniques influence SMS spam classification accuracy across machine learning models.

1.4. Objectives and Scope of the Study

The primary objective of this study is to analyze the impact of various text preprocessing techniques on the accuracy of SMS spam classification systems. Specifically, the study aims to evaluate individual and combined preprocessing methods and assess their effects on common performance metrics. The scope is limited to SMS spam datasets and focuses on traditional machine learning classifiers and standard NLP preprocessing approaches, providing practical guidelines for designing effective and efficient SMS spam detection pipelines.

2. Overview of SMS Spam Classification

2.1. Characteristics of SMS Data

SMS data possess unique characteristics that distinguish them from other text sources. Messages are typically short in length, which limits contextual information and increases data sparsity. They often contain noise in the form of misspellings, abbreviations, slang, emoticons, URLs, phone numbers, and special characters. Additionally, SMS messages are highly informal, lacking consistent grammar and punctuation. These characteristics make SMS spam classification challenging, as models must learn discriminative patterns from minimal and irregular textual content.

2.2. Common Machine Learning and Deep Learning Approaches

SMS spam classification has traditionally relied on machine learning algorithms such as Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees, and Random Forests. These models are commonly paired with bag-of-words or term frequency–inverse document frequency (TF-IDF) feature representations. More recently, deep learning approaches have gained prominence, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models. Deep learning models can automatically learn hierarchical and contextual features from text, often achieving higher accuracy, but they typically require larger datasets and greater computational resources.

2.3. Role of Feature Representation in Classification Accuracy

Feature representation is a critical factor influencing the accuracy of SMS spam classification systems. Effective feature representations capture meaningful patterns that distinguish spam from legitimate messages while minimizing noise and redundancy. Traditional representations such as bag-of-words and TF-IDF depend heavily on preprocessing quality, as vocabulary size and term relevance directly affect model learning. In contrast, word embeddings and contextual representations used in deep learning models encode semantic relationships between words. Regardless of the approach, the choice of feature representation, in combination with appropriate preprocessing techniques, plays a decisive role in determining classification performance.

3. Text Preprocessing in Natural Language Processing

3.1. Definition and Purpose of Text Preprocessing

Text preprocessing refers to a set of techniques used to transform raw textual data into a clean, structured, and machine-readable format suitable for natural language processing (NLP) tasks. Its primary purpose is to reduce noise, standardize text, and extract meaningful features that enhance the performance of computational models. Common preprocessing steps include text normalization, tokenization, stop-word removal, stemming, lemmatization, and handling of special characters or numbers. By preparing text data effectively, preprocessing improves model learning, reduces computational complexity, and increases the accuracy of tasks such as classification, sentiment analysis, and information retrieval.

3.2. Challenges Specific to SMS Text Preprocessing

SMS data pose unique challenges that make preprocessing particularly critical and complex:

- Short and Sparse Text: SMS messages are typically very brief, limiting the context available for feature extraction and increasing the difficulty of distinguishing spam from ham.
- Informal Language and Abbreviations: Users frequently employ slang, shorthand, or phonetic spellings, requiring normalization techniques to standardize text.
- Misspellings and Typos: Frequent errors in spelling can fragment word representations, reducing the effectiveness of models that rely on exact matches or vocabulary-based features.
- Special Characters and Emojis: SMS often contain emojis, symbols, or punctuation used to convey meaning, which can complicate tokenization and feature extraction.
- URLs, Phone Numbers, and Alphanumeric Codes: Spam messages often include links, codes, or numbers that are informative but must be handled carefully during preprocessing.
- Class Imbalance Sensitivity: Spam messages are usually fewer than ham messages, making preprocessing choices critical to preserving discriminative information without exacerbating imbalance.

Addressing these challenges requires careful selection and combination of preprocessing techniques to maximize classification accuracy while minimizing information loss.

4. Common Text Preprocessing Techniques

4.1. Text Normalization (Lowercasing, Punctuation Removal)

Text normalization standardizes raw text to reduce variability caused by inconsistent formatting. Lowercasing converts all characters to a uniform case, preventing duplicate representations of the same word (e.g., “Free” and “free”). Punctuation removal eliminates non-alphanumeric characters that may not contribute meaningful information. In SMS spam classification, normalization helps reduce vocabulary size and noise, although excessive removal of punctuation may discard useful cues such as repeated symbols often used in spam messages.

4.1.1. Tokenization

Tokenization is the process of splitting text into smaller units, typically words or subwords, known as tokens. For SMS data, tokenization must handle irregular spacing, emojis, URLs, and special symbols. Effective tokenization enables accurate feature extraction by ensuring that meaningful text components are correctly identified and represented in the model.

4.1.2. Stop-Word Removal

Stop-word removal eliminates commonly occurring words (e.g., “is,” “the,” “and”) that carry limited semantic value for classification. Removing stop words can reduce dimensionality and improve computational efficiency. However, in short SMS messages, aggressive stop-word removal may eliminate contextually important words, potentially reducing classification accuracy.

4.1.3. Stemming

Stemming reduces words to their root or base form by removing suffixes (e.g., “winning,” “winner” → “win”). This process helps consolidate word variants and reduce vocabulary size. While stemming is computationally efficient, it can produce non-linguistic root forms, which may negatively affect interpretability and, in some cases, model performance.

4.1.4. Lemmatization

Lemmatization maps words to their dictionary base form (lemma) using linguistic knowledge (e.g., “running” → “run”). Compared to stemming, lemmatization preserves semantic meaning and produces valid words. Although more computationally expensive, it often results in better feature quality for SMS spam classification tasks.

4.2. Handling Numbers, Symbols, and Special Characters

SMS spam messages frequently include phone numbers, currency symbols, URLs, and promotional codes. Preprocessing may involve removing, masking, or replacing these elements with placeholder tokens. Proper handling ensures that informative patterns, such as the presence of URLs or monetary values, are retained without introducing unnecessary noise.

4.3. Dealing with Slang, Abbreviations, and Misspellings

Informal language is common in SMS data, including slang, abbreviations, and intentional misspellings. Techniques such as slang dictionaries, abbreviation expansion, and spelling correction can improve text consistency and feature representation. However, these approaches must be applied carefully to avoid altering meaningful spam indicators or increasing preprocessing complexity.

5. Advanced and SMS-Specific Preprocessing Methods

5.1. URL, Email, and Phone Number Normalization

SMS spam messages frequently contain URLs, email addresses, and phone numbers as key indicators of malicious or promotional intent. Instead of removing these elements entirely, normalization techniques replace them with standardized placeholder tokens (e.g., <URL>, <EMAIL>, <PHONE>). This approach preserves their discriminative value while reducing vocabulary fragmentation caused by unique or randomly generated strings, thereby improving model generalization.

5.2. Emoji and Emoticon Handling

Emojis and emoticons are commonly used in SMS messages to express emotions or attract attention, particularly in spam messages. Preprocessing strategies include removing them, converting them into textual descriptions, or mapping them to sentiment-based tokens. Proper handling of emojis can help retain emotional or promotional cues that may contribute to improved classification accuracy.

5.3. Spell Correction and Text Expansion

Spell correction addresses typographical errors and intentional misspellings designed to bypass spam filters. Automated spelling correction tools can normalize word forms and reduce vocabulary sparsity. Text expansion techniques, such as converting abbreviations and shorthand (e.g., “u” to “you,” “msg” to “message”), further enhance text clarity. While beneficial, these methods must be applied cautiously to avoid altering meaningful patterns specific to spam content.

5.4. Handling Class Imbalance During Preprocessing

SMS spam datasets are often imbalanced, with legitimate messages significantly outnumbering spam messages. Preprocessing can play a role in mitigating this issue through techniques such as targeted data augmentation for minority classes, selective feature weighting, or careful preservation of rare but informative tokens. Addressing class imbalance during preprocessing helps prevent model bias toward the majority class and supports more robust spam detection performance.

6. Impact of Individual Preprocessing Techniques on Classification Accuracy

6.1. Effects on Traditional Machine Learning Models

Traditional machine learning models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression are highly sensitive to text preprocessing choices because they rely on explicit feature representations like bag-of-words and TF-IDF. Techniques such as lowercasing, stop-word removal, and stemming generally improve classification accuracy by reducing vocabulary size and noise. Tokenization quality directly affects feature consistency, while normalization of URLs and numbers often enhances spam detection performance. However, aggressive preprocessing such as excessive stop-word removal or stemming can remove discriminative terms, leading to reduced model accuracy, particularly for short SMS messages.

6.2. Effects on Deep Learning Models

Deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures, are comparatively more robust to raw and noisy text. These models can learn contextual and semantic patterns directly from sequences of words or subwords. As a result, minimal preprocessing such as basic normalization and tokenization is often sufficient. Techniques like lemmatization or stop-word removal may have limited impact and, in some cases, can negatively affect performance by altering word order or semantic cues. Nevertheless, normalization of URLs, emojis, and special tokens can still contribute positively by providing consistent input representations.

6.3. Trade-Offs Between Noise Reduction and Information Loss

While preprocessing aims to reduce noise and improve model efficiency, it introduces a critical trade-off between simplifying text and preserving meaningful information. Excessive preprocessing may discard subtle indicators of spam, such as repeated symbols, informal language, or specific keyword patterns. Conversely, insufficient preprocessing may leave irrelevant noise that degrades model learning. Achieving optimal classification accuracy therefore requires a balanced preprocessing strategy that minimizes noise while retaining features essential for distinguishing spam from legitimate SMS messages.

7. Comparative Analysis of Preprocessing Pipelines

7.1. Minimal vs. Extensive Preprocessing

Preprocessing pipelines can range from minimal approaches, involving basic normalization and tokenization, to extensive pipelines that include stop-word removal, stemming or lemmatization, spell correction, and SMS-specific normalization. Minimal preprocessing often preserves more of the original message structure and semantic cues, which can be beneficial for deep learning models. In contrast, extensive preprocessing tends to reduce noise and dimensionality, making it more suitable for traditional machine learning models. Comparative analysis shows that while extensive preprocessing can improve accuracy in some cases, it may also introduce information loss, particularly in short and informal SMS texts.

7.2. Model-Dependent Preprocessing Requirements

Different classification models exhibit varying sensitivity to preprocessing techniques. Traditional machine learning models typically require more aggressive preprocessing to optimize feature quality and reduce sparsity. Deep learning models, especially transformer-based architectures, can tolerate raw text and rely less on manual preprocessing due to their ability to learn contextual representations. As a result, the effectiveness of a preprocessing pipeline is strongly model-dependent, and no single pipeline performs optimally across all classifiers.

7.3. Performance Comparison Using Evaluation Metrics

The effectiveness of preprocessing pipelines is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. Comparative results indicate that preprocessing choices influence not only overall accuracy but also class-specific performance, particularly recall for spam detection. Balanced evaluation across multiple metrics is essential to identify preprocessing strategies that achieve robust and reliable SMS spam classification o

8. Evaluation Metrics and Experimental Setup

8.1. Accuracy, Precision, Recall, and F1-Score

To assess the effectiveness of SMS spam classification models under different preprocessing techniques, standard evaluation metrics are employed. Accuracy measures the overall proportion of correctly classified messages but may be misleading in imbalanced datasets**.** Precision evaluates the proportion of correctly identified spam messages among all messages classified as spam, reflecting the model's ability to avoid false positives. Recall measures the proportion of actual spam messages correctly detected, indicating the effectiveness of spam coverage. The F1-score**,** as the harmonic mean of precision and recall, provides a balanced assessment of model performance, particularly in the presence of class imbalance.

8.2. Dataset Description and Preprocessing Configurations

The experimental evaluation is conducted using benchmark SMS spam datasets containing labeled spam and legitimate messages. The datasets typically exhibit class imbalance, with legitimate messages forming the majority class. Multiple preprocessing configurations are designed to analyze their impact on classification performance, ranging from minimal preprocessing (lowercasing and tokenization) to extensive pipelines incorporating stop-word removal, stemming or lemmatization, normalization of URLs and numbers, and SMS-specific text handling. Each configuration is applied consistently across models to ensure fair comparison.

8.3. Experimental Methodology

The experimental methodology involves splitting the dataset into training and testing subsets using standard validation techniques. Classification models are trained separately under each preprocessing configuration, and their performance is evaluated using the selected metrics. Comparative analysis is then performed to identify trends, strengths, and limitations associated with different preprocessing strategies. This systematic approach ensures reproducibility and provides reliable insights into the relationship between text preprocessing techniques and SMS spam classification accuracy.

9. Challenges and Limitations

9.1. Over-Preprocessing and Semantic Distortion

One of the primary challenges in SMS spam classification is over-preprocessing, where excessive cleaning or normalization removes meaningful semantic information. Techniques such as aggressive stop-word removal, stemming, or symbol deletion can distort message intent, especially in short SMS texts where each token carries significant meaning. This semantic distortion may reduce the model's ability to capture subtle spam indicators, ultimately lowering classification accuracy.

9.2. Language and Domain Dependency

Preprocessing techniques are often language- and domain-specific, limiting their generalizability. Methods optimized for English SMS data may not perform effectively on messages in other languages or mixed-language contexts. Additionally, spam content varies across regions and domains, requiring adaptation of preprocessing rules, slang dictionaries, and normalization strategies. This dependency poses challenges for building universally robust SMS spam classification systems.

9.3. Computational Cost and Scalability

Advanced preprocessing methods such as spell correction, lemmatization, and text expansion increase computational complexity and processing time. In large-scale or real-time SMS filtering systems, these costs can impact scalability and deployment feasibility. Balancing preprocessing sophistication with computational efficiency remains a key limitation, particularly in resource-constrained environments where rapid message classification is required.

10. Implications for SMS Spam Detection Systems

10.1. Best Practices for Preprocessing Selection

The findings of SMS spam classification studies highlight the importance of selecting preprocessing techniques that align with the chosen classification model and application context. For traditional machine learning models, structured and moderately extensive preprocessing such as normalization, tokenization, and controlled stemming or lemmatization tends to yield better performance. For deep learning models, minimal but consistent preprocessing is often sufficient, with emphasis on preserving contextual and semantic information. In all cases, preprocessing pipelines should be empirically evaluated rather than assumed, as their impact varies across datasets and models.

10.2. Balancing Accuracy, Efficiency, and Robustness

An effective SMS spam detection system must balance high classification accuracy with computational efficiency and robustness to evolving spam patterns. Overly complex preprocessing pipelines may improve accuracy marginally but at the cost of increased latency and reduced scalability. Conversely, insufficient preprocessing may lead to noisy inputs and unstable predictions. A balanced approach that combines essential noise reduction with preservation of discriminative features supports reliable performance across diverse message types and operating conditions.

10.3. Practical Deployment Considerations

In real-world deployments, SMS spam detection systems must operate under constraints such as real-time processing, limited computational resources, and dynamic spam behavior. Preprocessing techniques should therefore be lightweight, adaptable, and easy to update as new spam patterns emerge. Additionally, system designers should consider maintainability, language support, and integration with existing communication infrastructures. These practical considerations ensure that preprocessing strategies contribute effectively to the long-term reliability and usability of SMS spam detection systems.

11. Future Research Directions

11.1. Adaptive and Automated Preprocessing Techniques

Future research can focus on developing adaptive and automated preprocessing methods that dynamically adjust to data characteristics and model requirements. Instead of relying on fixed preprocessing pipelines, learning-based or data-driven approaches can identify optimal preprocessing strategies based on message content, noise level, or evolving spam patterns. Such adaptive techniques have the potential to reduce manual tuning and improve long-term classification performance.

11.2. Multilingual and Cross-Domain Preprocessing Strategies

As SMS communication increasingly spans multiple languages and domains, there is a growing need for preprocessing strategies that generalize beyond single-language or domain-specific settings. Future studies may explore language-agnostic preprocessing methods, cross-lingual normalization techniques, and transfer learning approaches that enable effective spam detection across diverse linguistic contexts. Addressing code-switching and mixed-language SMS content is also a critical research direction.

11.3. Integration with Contextual Embeddings

The integration of preprocessing techniques with contextual word and sentence embeddings represents a promising area for future research. Modern embedding models capture semantic and contextual information that may reduce the need for aggressive preprocessing. Investigating how minimal or selective preprocessing interacts with contextual embeddings can lead to more efficient and accurate SMS spam classification systems, particularly when combined with transformer-based architectures.

12. Conclusion

12.1. Summary of Key Findings

This study examined the role of text preprocessing in SMS spam classification and analyzed how different preprocessing techniques influence model performance. The findings show that preprocessing significantly affects feature representation and classification outcomes, particularly for short and noisy SMS data. Traditional machine learning models benefit from structured and carefully designed preprocessing pipelines, while deep learning models demonstrate greater robustness to raw text and require less aggressive preprocessing.

12.2. Overall Impact of Preprocessing on SMS Spam Classification Accuracy

Overall, text preprocessing has a substantial impact on SMS spam classification accuracy by reducing noise, controlling vocabulary size, and enhancing discriminative feature extraction. However, the results also highlight that excessive preprocessing can lead to semantic distortion and information loss, negatively affecting performance. The effectiveness of preprocessing is therefore model-dependent, dataset-specific, and closely tied to the nature of SMS content.

12.3. Recommendations for Researchers and Practitioners

Researchers are encouraged to systematically evaluate preprocessing techniques rather than relying on standard or assumed pipelines. Future studies should explore adaptive, multilingual, and context-aware preprocessing approaches to improve generalization. Practitioners should select preprocessing strategies that balance accuracy, efficiency, and scalability, prioritizing lightweight and maintainable solutions for real-world SMS spam detection systems.

Reference

[1] Narra, B., Buddula, D. V. K. R., Patchipulusu, H., Vattikonda, N., Gupta, A., & Polu, A. R. (2024). The integration of artificial intelligence in software development: Trends, tools, and future prospects. Available at SSRN 5596472.

[2] Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Vattikonda, N. (2024). Leveraging deep learning models for intrusion detection systems for secure networks. Journal of Computer Science and Technology Studies, 6(2), 199-208.

[3] Achuthananda, R. P., Bhumeka, N., Dheeraj Varun Kumar, R. B., Hari Hara, S. P., & Navya, V. (2024). Evaluating machine learning approaches for personalized movie recommendations: A comprehensive analysis. J Contemp Edu Theo Artific Intel: JCETAI-115.

[4] Polu, A. R., Narra, B., Buddula, D. V. K. R., Hara, H., Patchipulusu, S., Vattikonda, N., & Gupta, A. K. Analyzing The Role of Analytics in Insurance Risk Management: A Systematic Review of Process Improvement and Business Agility.

[5] Tamilmani, V., Maniar, V., Singh, A. A., Kothamaram, R. R., Rajendran, D., & Namburi, V. D. (2024). A Review of Cyber Threat Detection in Software-Defined and Virtualized Networking Infrastructures. International Journal of Technology, Management and Humanities, 10(04), 136-146.

[6] Kothamaram, R. R., Rajendran, D., Namburi, V. D., Tamilmani, V., Maniar, V., & Singh, A. A. S. Predictive Analytics for Customer Retention in Telecommunications Using ML Techniques.

[7] Singh, A. A. S., Kothamaram, R. R., Rajendran, D., Deepak, V., Namburi, V. T., & Maniar, V. A Review on Model-Driven Development with a Focus on Microsoft PowerApps.

[8] Bitkuri, V., Kendyala, R., Kurma, J., Mamidala, J. V., Attipalli, A., & Enokkaren, S. J. (2024). A Survey on Blockchain-Enabled ERP Systems for Secure Supply Chain Processes and Cloud Integration. International Journal of Technology, Management and Humanities, 10(04), 126-135.

[9] Waditwar, P. (2024) AI for Bathsheba Syndrome: Ethical Implications and Preventative Strategies. Open Journal of Leadership, 13, 321-341. doi: 10.4236/ojl.2024.133020

[10] Mamidala, J. V., Bitkuri, V., Attipalli, A., Kendyala, R., Kurma, J., & Enokkaren, S. J. (2024). Machine Learning Approaches to Salary Prediction in Human Resource Payroll Systems. Journal of Computer Science and Technology Studies, 6(5), 341-349.

[11] Attipalli, A., Kendyala, R., Kurma, J., Mamidala, J. V., Bitkuri, V., & Enokkaren, S. J. Privacy Preservation in the Cloud: A Comprehensive Review of Encryption and Anonymization Methods. International Journal of Multidisciplinary on Science and Management IJMSM, 1(1).

[12] Enokkaren, S. J., Kendyala, R., Kurma, J., Mamidala, J. V., Bitkuri, V., & Attipalli, A. Artificial Intelligence (AI)-Based Advance Models for Proactive Payroll Fraud Detection and Prevention.

[13] Gangineni, V. N., Tyagadurgam, M. S. V., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2024). AI-Powered Cybersecurity Risk Scoring for Financial Institutions Using Machine Learning Techniques (Approved by ICITET 2024). Journal of Artificial Intelligence & Cloud Computing.

[14] Bitkuri, V., Kendyala, R., Kurma, J., Mamidala, J. V., Enokkaren, S. J., & Attipalli, A. (2022). Empowering Cloud Security with Artificial Intelligence: Detecting Threats Using Advanced Machine learning Technologies. International Journal of AI, BigData, Computational and Management Studies, 3(4), 49-59.

[15] Attipalli, A., Mamidala, J. V., KURMA, J., Bitkuri, V., Kendyala, R., & Enokkaren, S. (2022). Towards the Efficient Management of Cloud Resource Allocation: A Framework Based on Machine Learning. Available at SSRN 5741265.

[16] Enokkaren, S. J., Attipalli, A., Bitkuri, V., Kendyala, R., Kurma, J., & Mamidala, J. V. (2022). A Deep-Review based on Predictive Machine Learning Models in Cloud Frameworks for the Performance Management. Universal Library of Engineering Technology, (Issue).

[17] Kurma, J., Mamidala, J. V., Attipalli, A., Enokkaren, S. J., Bitkuri, V., & Kendyala, R. (2022). A Review of Security, Compliance, and Governance Challenges in Cloud-Native Middleware and Enterprise Systems. International Journal of Research and Applied Innovations, 5(1), 6434-6443.

[18] Attipalli, A., Enokkaren, S., KURMA, J., Mamidala, J. V., Kendyala, R., & BITKURI, V. (2022). A Deep-Review based on Predictive Machine Learning Models in Cloud Frameworks for the Performance Management. Available at SSRN 5741282.

[19] Bitkuri, V., Kendyala, R., Kurma, J., Mamidala, J. V., Enokkaren, S. J., & Attipalli, A. (2022). Empowering Cloud Security with Artificial Intelligence: Detecting Threats Using Advanced Machine learning Technologies. International Journal of AI, BigData, Computational and Management Studies, 3(4), 49-59.

[20] Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., & Bhumireddy, J. R. (2022). Leveraging big datasets for machine learning-based anomaly detection in cybersecurity network traffic. Available at SSRN 5538121.

[21] Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., & Nandiraju, S. K. K. (2022). Efficient machine learning approaches for intrusion identification of DDoS attacks in cloud networks. Available at SSRN 5515262.

[22] Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., & Bhumireddy, J. R. (2022). Leveraging big datasets for machine learning-based anomaly detection in cybersecurity network traffic. Available at SSRN 5538121.

[23] Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. J Contemp Edu Theo Artific Intel: JCETAI/101.

[24] Namburi, V. D., Singh, A. A. S., Maniar, V., Tamilmani, V., Kothamaram, R. R., & Rajendran, D. (2023). Intelligent Network Traffic Identification Based on Advanced Machine Learning Approaches. International Journal of Emerging Trends in Computer Science and Information Technology, 4(4), 118-128.

[25] Rajendran, D., Maniar, V., Tamilmani, V., Namburi, V. D., Singh, A. A. S., & Kothamaram, R. R. (2023). CNN-LSTM Hybrid Architecture for Accurate Network Intrusion Detection for Cybersecurity. Journal Of Engineering And Computer Sciences, 2(11), 1-13.

[26] Kothamaram, R. R., Rajendran, D., Namburi, V. D., Tamilmani, V., Singh, A. A., & Maniar, V. (2023). Exploring the Influence of ERP-Supported Business Intelligence on Customer Relationship Management Strategies. International Journal of Technology, Management and Humanities, 9(04), 179-191.

[27] Singh, A. A. S. S., Mania, V., Kothamaram, R. R., Rajendran, D., Namburi, V. D. N., & Tamilmani, V. (2023). Exploration of Java-Based Big Data Frameworks: Architecture, Challenges, and Opportunities.Journal of Artificial Intelligence & Cloud Computing,2(4), 1-8.

[28] Waditwar, P. (2024) The Intersection of Strategic Sourcing and Artificial Intelligence: A Paradigm Shift for Modern Organizations. Open Journal of Business and Management, 12, 4073-4085. doi: 10.4236/ojbm.2024.126204

[29] Almeida, T. A., Gómez Hidalgo, J. M., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. Proceedings of the 11th ACM Symposium on Document Engineering, 259–262.

[30] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289–1305.

[31] Gómez Hidalgo, J. M., Bringas, G. C., Sánz, E. P., & García, F. C. (2006). Content based SMS spam filtering. Proceedings of the 2006 ACM Symposium on Document Engineering, 107–114.

[32] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning, 137–142. Springer.

[33] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[34] Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning, 133–142.

[35] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. Learning for Text Categorization: Papers from the 1998 Workshop, 55–62.

[36] Chundru, S. K., Vikram, M. S., Naidu, V., Pabbineedi, S., Kakani, A. B., & Nandiraju, S. K. K. Analyzing and Predicting Anaemia with Advanced Machine Learning Techniques with Comparative Analysis.

[37] Polam, R. M., Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Vangala, S. R. (2025). Advanced Machine Learning for Robust Botnet Attack Detection in Evolving Threat Landscapes. Available at SSRN 5515384.

[38] Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Vangala, S. R., & Polam, R. M. (2025). Data-Driven Detection of Network Threats using Advanced Machine Learning Techniques for Cybersecurity. Available at SSRN 5515400.

[39] Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2025). Effectiveness of Deep Learning Algorithms in Phishing

Attack Detection for Cybersecurity Frameworks. Available at SSRN 5515385.

[40] Vanaparthi, N. R. (2025). Why digital transformation in fintech requires mainframe modernization: A cost-benefit analysis. International Journal of Science and Research Archive, 14(1), 1052–1062. https://doi.org/10.30574/ijsra.2025.14.1.0161

[41] Kamarthapu, B., Penmetsa, M., Vangala, S. R., & Polam, R. M. (2025). Effectiveness of Deep Learning Algorithms in Phishing Attack Detection for Cybersecurity Frameworks. Available at SSRN 5571241.

[42] Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2025). Leveraging NLP and Sentiment Analysis for ML-Based Fake News Detection with Big Data. Available at SSRN 5515418.

[43] Gangineni, V. N., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. SV, & Pabbineedi, S.(2025). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce.

[44] Prajkta Waditwar. Quantum-Enhanced Travel Procurement: Hybrid Quantum–Classical Optimization for Enterprise Travel Management. World Journal of Advanced Engineering Technology and Sciences, 2025, 17(03), 375-386. Article DOI: https://doi.org/10.30574/wjaets.2025.17.3.1572.

[45] Vanaparthi, N. R. (2025). Regulatory compliance in the digital age: How mainframe modernization can support financial institutions. International Journal of Research in Computer Applications and Information Technology, 8(1), 383–396. https://doi.org/10.34218/IJRCAIT_08_01_033

[46] Waditwar, P. (2025) AI-Driven Procurement in Ayurveda and Ayurvedic Medicines & Treatments. Open Journal of Business and Management, 13, 1854-1879. doi: 10.4236/ojbm.2025.133096

[47] Vanaparthi, N. R. (2025). The roadmap to mainframe modernization: Bridging legacy systems with the cloud. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(1), 125–133. https://doi.org/10.32628/CSEIT25111214

[48] Prabakar, D., Iskandarova, N., Iskandarova, N., Kalla, D., Kulimova, K., & Parmar, D. (2025, May). Dynamic Resource Allocation in Cloud Computing Environments Using Hybrid Swarm Intelligence Algorithms. In 2025 International Conference on Networks and Cryptology (NETCRYPT) (pp. 882-886). IEEE.

[49] Nagaraju, S., Johri, P., Putta, P., Kalla, D., Polvanov, S., & Patel, N. V. (2025, May). Smart Routing in Urban Wireless Ad Hoc Networks Using Graph Attention Network-Based Decision Models. In 2025 International Conference on Networks and Cryptology (NETCRYPT) (pp. 212-216). IEEE.

[50] Kalla, D., Mohammed, A. S., Boddapati, V. N., Jiwani, N., & Kiruthiga, T. (2024, November). Investigating the Impact of Heuristic Algorithms on Cyberthreat Detection. In 2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (Vol. 1, pp. 450-455). IEEE.

[51] Vadisetty, R., Polamarasetti, A., & Kalla, D. (2025, February). Automated AI-Driven Phishing Detection and Countermeasures for Zero-Day Phishing Attacks. In International Ethical Hacking Conference (pp. 285-303). Singapore: Springer Nature Singapore.

[52] Nagrath, P., Saini, I., Zeeshan, M., Komal, Komal, & Kalla, D. (2025, June). Predicting Mental Health Disorders with Variational Autoencoders. In International Conference on Data Analytics & Management (pp. 38-51). Cham: Springer Nature Switzerland.

[53] Polam, R. M., Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Vangala, S. R. (2025). Advanced Machine Learning for Robust Botnet Attack Detection in Evolving Threat Landscapes. Available at SSRN 5515384.

[54] Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Vangala, S. R., & Polam, R. M. (2025). Data-Driven Detection of Network Threats using Advanced Machine Learning Techniques for Cybersecurity. Available at SSRN 5515400.

[55] Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2025). Effectiveness of Deep Learning Algorithms in Phishing Attack Detection for Cybersecurity Frameworks. Available at SSRN 5515385.

[56] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2025). Towards Early Forecast of Diabetes Mellitus via Machine Learning Systems in Healthcare. European Journal of Technology, 9(1), 35-50.

[57] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2025). Predictive Modeling for Property Insurance Premium Estimation Using Machine Learning Algorithms. Available at SSRN 5515382.

[58] Nandiraju, S. K. K., & Chundru, S. K. Enhancing Cybersecurity: Zero-Day.

[59] Prajkta Waditwar. Agentic AI and sustainable procurement: Rethinking anti-corrosion strategies in oil and gas. World Journal of Advanced Research and Reviews, 2025, 27(03), 1591-1598. Article DOI: https://doi.org/10.30574/wjarr.2025.27.3.3298.

[60] Vadisetty, R., Polamarasetti, A., Varadarajan, V., Kalla, D., & Ramanathan, G. K. (2025, May). Cyber Warfare and AI Agents: Strengthening National Security Against Advanced Persistent Threats (APTs). In International Conference on Intelligence-Based Transformations of Technology and Business Trends (pp. 578-587). Cham: Springer Nature Switzerland.