Reducing Human-Induced Label Bias in SMS Spam with Context-Enhanced Clustering (CEC)

Gerard Shu Fuhnwi*, Ann Marie Reinhold*†, Clemente Izurieta*†‡,

*Montana State University, Bozeman MT, USA
†Pacific Northwest National Laboratory, Richland WA, USA
‡Idaho National Laboratory, Idaho Falls ID, USA

***Abstract*—Short Message Service (SMS) is a widely used text messaging feature available on both basic and smartphones, making SMS spam detection a critical task. Supervised machine learning approaches often face challenges in this domain due to their dependence on manually crafted features, such as keyword detection, which can result in simplistic patterns and misclassification of more complex messages. Furthermore, these models can exacerbate human-induced bias if the training data include inconsistent labeling or subjective interpretations, leading to unfair treatment of specific keywords or contexts.**

We propose a Context-Enhanced Clustering (CEC) approach to address these challenges by leveraging contextual metadata, adaptive thresholding, and modified similarity measures for clustering. We evaluate our approach using the English SMS spam dataset source from UC Irvine's Machine Learning Repository. CEC identifies representative samples from the SMS dataset to fine-tune LLMs such as ChatGPT-4, improving the robustness and fairness of spam classification. Our approach outperforms traditional clustering techniques such as K-means and DBSCAN in mitigating bias, as demonstrated through experiments measuring a balanced accuracy of 85% and a treatment equality difference (TED) of precisely zero. When used to identify representative samples to fine-tune ChatGPT-4, the CEC achieves a balanced accuracy of 98%, an equal opportunity of difference (EOD), and a treatment equality difference (TED) of zero. These results significantly reduce human-induced bias while maintaining high classification accuracy.

***Index Terms*—CEC, SMS, Bias, Mitigation, EOD, TED, SPD, LLM**

I. Introduction

Short Message Service (SMS) is not merely a mode of communication but a global phenomenon. Despite the increasing dominance of Internet messaging platforms, SMS is one of the most ubiquitous and reliable communication methods. In 2023, there were 6.89 billion smartphone users worldwide, projected to increase to 7.86 billion by 2028. SMS remains highly effective, with 95% of text messages read and responded to within three minutes after receiving. In the United States alone, 97% of adults own mobile phones (85% of which are smartphones), and the SMS market is projected to reach a value of 12.6 billion dollars by 2025, experiencing a significant growth compound annual rate of 20.3% [1]. However,the simplicity and widespread availability of mobile devices have made SMS a significant target for spam messages, leading to fraud, phishing, and identity theft [1]. In 2022, a survey in the United States revealed that more than 225 billion spam texts were sent, a staggering 157% increase from 87.8 billion in 2021. These scams resulted in an estimated 20 billion dollars in financial losses[2].

Furthermore, spam significantly burdens network capacity and data storage, further compounding its negative impacts [2]. As SMS usage increases, the demand for effective and robust spam detection systems becomes increasingly critical. Addressing this challenge is essential to ensure user safety, maintaining network efficiency, and preserving SMS's trust and reliability as a communication platform. Existing spam detection methods often rely on manually labeled datasets, which introduce human biases arising from inconsistencies in the way messages are labeled, affecting classification fairness and accuracy [3], [4]. Although clustering techniques such as K-means [5] and DBSCAN [6] are widely used for pre-processing and semisupervised learning, they fail to address these biases due to static configurations and limited adaptability to contextual features.

Human-induced bias refers to systematic patterns of unfairness introduced into data due to human decisions, perspectives, or behaviors that can result in discriminatory outcomes. Spam detection plays a crucial role in ensuring the security and reliability of SMS communications while minimizing economic losses for organizations [7]. However, its effectiveness is often compromised by the human-induced label bias present in the training datasets [4]. This bias can arise from subjective labeling, inconsistent interpretations, annotator fatigue, limited domain expertise, or a lack of transparency in the labeling process. Such biases introduce inconsistencies in the training data [8], ultimately distorting the models developed for spam classification.

Several clustering techniques have been proposed for SMS spam detection, offering a combination of methods to enhance effectiveness [9]–[13]. However, in most existing work, incorporating contextual metadata, such as spam-related keywords and fairness, still needs to be addressed, limiting their applicability in real-world scenarios. This motivates the development of a fair, efficient, and context-aware clustering approach, such as Context-Enhanced

[1]https://www.smscomparison.com/sms-statistics/

[2]https://www.robokiller.com/robokiller-2022-phone-scam-report

Clustering for SMS spam detection, which incorporates adaptive techniques and contextual metadata to improve clustering quality and reduce human-induced labeling bias. Our research specifically addresses the following questions:

- **RQ1:** How effective is the context-driven clustering approach in reducing human-induced label bias compared to traditional clustering techniques such as DBSCAN and K-means in SMS spam detection?
- **RQ2:** Can a context-driven clustering technique effectively automate the selection of representative samples to generate prompts to fine-tune large-language models in SMS spam detection?

The main contributions of our approach are as follows:

- We introduce a novel context-driven clustering framework that incorporates contextual metadata, modified cosine similarity, and adaptive thresholding to improve clustering quality in SMS spam detection.
- We automate the selection of representative samples from clusters, enabling efficient and unbiased prompt generation for LLM fine-tuning and eliminating manual effort.
- We demonstrate significant reductions in human-induced label bias and improved fairness in SMS spam detection using metrics such as equal opportunity, statistical parity, and treatment equality differences while maintaining high classification performance.

The rest of this paper is organized as follows: Section II discusses related work. Section III introduces the Context-Enhanced SMS Clustering (CESC) framework, detailing the integration of contextual metadata, modified cosine similarity, adaptive thresholding, and the selection of representative samples for fine-tuning LLMs. Section IV describes model performance evaluation, fairness metrics, and results addressing the research questions (RQs). Section V provides an in-depth discussion of the empirical results. Section VI presents a threat to the validity of our approach. Finally, Section VII concludes the paper and suggests potential directions for future research.

II. Related Work

Human-induced bias in data labeling remains a persistent challenge in machine learning, especially for tasks that require subjective interpretations, such as spam detection. When annotators label large datasets, their perspectives can introduce inconsistencies that adversely affect the performance of models trained on these data. This variability poses a considerable challenge in tasks such as SMS spam detection. Research, including the work by Fort et al. [14], highlights how human labelers' backgrounds and personal biases can shape their labeling decisions. Although several research studies have been conducted about data labeling in the SMS spam detection domain using clustering algorithms such as K-means [5] and DBSCAN [6], little research has been conducted to address human-induced biases due to the large number of short messages in SMS.

A contextual term appears to be a particular term in short-text messages such as "text", "free," etc. Therefore, special clustering techniques are needed to reduce human-induced label bias in SMS text messages for spam detection, since existing clustering methods have limitations in adapting to contextual terms in short text data such as SMS.

Specifically, with respect to SMS detection, Nagwani and Sharaff [9] investigated a bi-level text classification and clustering approach employing K-means to improve SMS spam filtering and thread identification. Their approach demonstrated that integrating clustering with classification techniques can effectively enhance spam detection but has limited adaptability to varying message densities.

Anjali et al. [10] introduced optical character recognition for image data using unsupervised and deep semi-supervised learning for the detection of SMS scams. This study employs a combination of K-means, Non-Negative Matrix Factorization, and Gaussian Mixture Models along with feature extraction techniques such as TF-IDF and PCA, with K-means feature extraction and vectorizer achieving a superior accuracy.

Hind and Rachid [11] explored unsupervised and supervised learning techniques, a hybrid model combining K-means and Naive Bayes, Random Forests, Logistic regression, and a Support Vector Machine for SMS spam detection, with K-means-SVM having outstanding precision. Similarly, Darshit [12] also combines clustering with the Support Vector Machine for SMS spam detection.

A comparative analysis by Songfeng et al. [15] evaluated the performance of K-means and DBSCAN on synthetic datasets. The study provided valuable insights into the strengths and limitations of each algorithm, informing their application in SMS spam detection. In a similar research work by Ahmad and Shilpa [13], they analyze four clustering algorithms, namely K-means, DBSCAN, HCA, and MDBCA, and compare their performance using different datasets.

These studies demonstrate the ability of clustering for text classification. Our study builds on these papers to extend the works of [9]–[12], [15] by using context-enhanced clustering to reduce human-induced labeling bias in SMS spam detection. We expand on previous approaches by addressing the evolving nature of spam messages and contextual features that are present in short SMS texts.

III. Approach

This section describes the Context-Enhanced Clustering (CEC) framework for mitigating human-induced labeling bias in SMS spam detection. CEC leverages contextual metadata, modified cosine similarity, and adaptive clustering to help create high-quality clusters of SMS messages as shown in Figure 1. Our CEC approach automatically selects

representative samples from the generated clusters, ensuring a diverse and unbiased selection of messages. These representative samples are then used to create prompts for fine-tuning a Lareg Language Model (LLM), enhancing its ability to classify SMS spam more fairly and accurately while mitigating human-induced labeling bias.

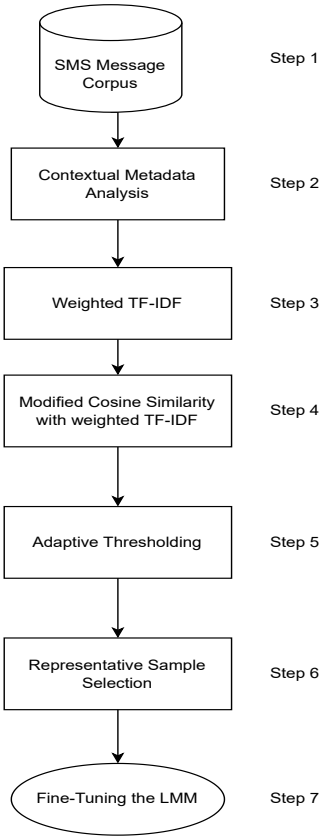

Fig. 1. Flow chart of the CEC approach. The rectangles represent the CEC processes, the cylinders represent stored data, the oval shape represents the final step, and the arrows indicate the flow direction.

A. SMS Spam Data (Step 1)

The SMS spam collection dataset from the UCI Machine Learning Repository comprises 5572 entries with two columns [16]. The first column specifies the message category, either spam (unsolicited message) or ham (regular message), while the second column contains the message text. Of the 5572 entries, 4825 are ham messages (86.6 %) and 747 spam messages (13.4 %).

B. Contextual Metadata Analysis (Step 2)

CEC begins by analyzing the contextual metadata of SMS messages to assign importance to terms that are more relevant to spam messages. Spam-related keywords ("free," "win," "offer") are identified based on domain knowledge and their frequency in spam-labeled messages. Each term t_{ij} in a message m_i is assigned a weight based on its relevance and frequency:

$$w_{ij} = \text{Relevance}(t_{ij}) \times \text{Frequency}(t_{ij}, M_{\text{spam}})$$

Where w_{ij} is the contextual weight for term t_{ij}, calculated based on term relevance and its frequency in spam, and M_{spam} denotes the subset of messages labeled as spam.

C. Weighted TF-IDF (Step 3)

Each SMS message is transformed into a weighted Term Frequency Inverse Document Frequency (TF-IDF) vector, emphasizing contextually essential terms. The weighted TF-IDF score for a term t_{ij} is computed as:

$$\text{Weighted TF-IDF}(t_{ij}) = \text{TF-IDF}(t_{ij}) \cdot w_{ij}$$

The weighted TF-IDF helps capture the importance of each term within a message, accounting for the contextual significance in spam detection.

D. Modified Cosine Similarity with Weighted TF-IDF (Step 4)

A modified cosine similarity metric is applied to measure similarity between SMS messages, enhancing traditional similarity computation by assigning a greater weight to spam-indicative keywords. Consider two messages $A = [a_1, a_2, \ldots, a_n]$ and $B = [b_1, b_2, \ldots, b_n]$. Let $W = [w_1, w_2, \ldots, w_n]$ represent the vector of contextual weights. Then, the modified cosine similarity is defined as:

$$\text{Modified Cosine Similarity}_{AB} = \frac{\sum_{k=1}^{n} w_k \cdot a_k \cdot b_k}{\sqrt{\sum_{k=1}^{n} w_k \cdot a_k^2} \cdot \sqrt{\sum_{k=1}^{n} w_k \cdot b_k^2}}$$

Where w_k is the contextual weight of term k, and a_k, b_k are the weighted term frequencies for messages A and B, respectively.

E. Adaptive Thresholding (Step 5)

To allow clusters to adapt dynamically, thresholds for grouping messages must be determined. For each message m_i, the local density is calculated using the average similarity of its k-nearest neighbors:

$$D_i = \frac{1}{k} \sum_{j \in \text{Neighbors}_k(i)} S_{ij}$$

where S_{ij} is the similarity between messages m_i and m_j. The thresholding for clustering is then calculated as:

$$\epsilon_i = \alpha \cdot D_i$$

The scaling factor α was tuned in the range $\alpha \in [0.3, 1.5]$, with optimal performance in $\alpha = 0.5$, where α adjusts the

sensitivity to the density of local messages. This adaptive thresholding enables tighter grouping in high-density regions to prevent overgeneralization and looser thresholds in low-density areas to capture rare spam patterns effectively.

F. Representative Sample Selection (Step 6)

When the adaptive clusters are formed based on local message density as described in step 5, representative samples are selected from these clusters to provide a balanced and unbiased subset for fine-tuning LLMs. These samples are selected based on the centrality score for a message x in the cluster C_i. Messages with the highest centrality score are selected as representative samples, ensuring they are most indicative of the cluster characteristics. This centrality score is calculated as the average similarity S_{xy} between x and all other messages y in the cluster C_i as follows:

$$R(x) = \frac{1}{|C_i|} \sum_{y \in C_i} S_{xy}$$

G. Fine-Tuning the LLM (Step 7)

The selected representative samples generate prompts for fine-tuning a pre-trained LLM, such as ChatGPT, for SMS spam detection. LLM fine-tuning trains the model to classify messages as spam or non-spam, using the selected samples as input. This process reduces human-induced label bias by relying on the contextually balanced generated prompts from the clusters, which helps improve both fairness and model performance in classification.

The pseudo-code of our approach is shown in Figure 2.

Algorithm 1 Context-Enhanced Clustering (CEC)

Require: SMS corpus M, spam-related keywords K, TF-IDF, LLM model

Preprocess messages and compute TF-IDF vectors
Assign contextual weights to terms in K
Compute weighted TF-IDF vectors
Calculate modified cosine similarity for all messages
Perform adaptive thresholding:
for each message m_i **do**
    Compute local density using k-nearest neighbors
    Calculate threshold $\epsilon_i = \alpha \times$ mean_density
end for
Perform adaptive clustering:
Initialize clusters as empty
for each pair of messages (m_i, m_j) **do**
    if $S_{ij} \geq \min(\epsilon_i, \epsilon_j)$ **then**
        Assign m_i and m_j to the same cluster
    end if
end for
for each cluster **do**
    Calculate centrality scores for messages
    Select representative samples with the highest scores
end for
Fine-tune LLM using representative samples as prompts
return Fine-tuned LLM model for spam detection

Fig. 2. Pseudocode for Context-Enhanced Clustering (CEC) Approach.[3]

[3]https://github.com/gshufuhnwi/CEC-Approach

IV. Results Evaluation

Here, we describe how we evaluated our approach and provide the results that answer our research questions.

A. Model Performance

The performance of our approach was evaluated using three performance metrics: precision (P), recall (R), and balanced accuracy (ACC). Precision (P) measures the proportion of correctly predicted positive (spam) examples relative to all examples classified as positive (spam):

$$P = \frac{TP}{TP + FP}$$

Recall is the proportion of positive instances (spam) correctly predicted out of the total number of actual positive examples:

$$R = \frac{TP}{TP + FN}$$

Balanced Accuracy (ACC) is the mean accuracy calculated across both the positive (spam) and negative (ham) classes steenhoek2024comprehensive:

$$ACC = \frac{\frac{correct_{negative}}{examples_{negative}} + \frac{correct_{positive}}{examples_{positive}}}{2}$$

B. Fairness Metric

Fairness refers to the ability of a spam detection model to classify spam and ham messages equitably without introducing biases that disproportionately affect specific message patterns, keywords, or context. We evaluated fairness using three metrics: statistical parity difference (SPD), equal opportunity difference (EOD), and treatment equality difference (TED) [18]. Statistical parity difference quantifies the difference in the likelihood of a message being classified as spam based on the presence of keywords such as "free" and "win."

$$\text{SPD} = \left| P(\text{Spam} \mid \text{"free"}) - P(\text{Spam} \mid \text{"win"}) \right|$$

where,

$$P(\text{Spam} \mid \text{"free"}) = \left| \frac{TP_{free} + FP_{free}}{TP_{free} + FP_{free} + FN_{free} + TN_{free}} \right|$$

and

$$P(\text{Spam} \mid \text{"win"}) = \left| \frac{TP_{win} + FP_{win}}{TP_{win} + FP_{win} + FN_{win} + TN_{win}} \right|$$

Equal opportunity difference (EOD) calculates the difference in true positive rates for spam messages containing keywords such as "free" and "text."

$$\text{EOD} = |P(\text{True Positive} \mid \text{"free"}) - P(\text{True Positive} \mid \text{"}text\text{"})|$$

where,

$$P(\text{True Positive} \mid \text{"free"}) = \frac{TP_{\text{free}}}{TP_{\text{free}} + FN_{\text{free}}}$$

and

$$P(\text{True Positive} \mid \text{"win"}) = \frac{TP_{\text{win}}}{TP_{\text{win}} + FN_{\text{win}}}$$

Treatment equality difference (TED) is satisfied if messages containing spam keywords like "free" and "win" have an equal ratio of false negatives (FN) and false positives (FP).

$$\text{TED} = \left|\left(\frac{\text{FN}}{\text{FP}}\right)_{\text{free}} - \left(\frac{\text{FN}}{\text{FP}}\right)_{\text{win}}\right|$$

SPD, EOD, and TED values near zero indicate that the algorithm is equally likely to make errors for each spam keyword, reflecting the fairness of its decision-making across keywords in spam messages. In this evaluation, the keywords "free" and "win" were selected based on their high contextual weights, representing the most influential terms associated with spam, making them ideal for the evaluation of fairness.

C. Addressing the Research Questions (RQs)

1) RQ1:: How effective is the context-driven clustering approach in reducing human-induced label bias compared to traditional clustering techniques such as DBSCAN and K-means in SMS spam detection?

To answer this research question, we applied our approach described in Section 3, which uses contextual metadata, modified cosine similarity, and adaptive clustering techniques to generate high quality clusters of SMS messages to help reduce human-induced label bias. Our approach demonstrated a balanced accuracy of 85%, recall of 68%, and precision of 100%, highlighting its effectiveness in accurately classifying spam and ham SMS messages. Furthermore, it maintained fairness with a treatment equality difference of 0, as shown in Table I. For comparison, K-means and DBSCAN were configured with key parameters set as follows: The K-means were run with the number of clusters = 20, while DBSCAN used eps = 0.9 and minimum samples = 2. DBSCAN, with similar precision and low accuracy, is better suited for scenarios where minimizing false positives is a priority but might fail to deliver consistent results across all messages. In contrast, K-means with low precision and low accuracy are unsuitable for applications where reliable and accurate message classification is essential. DBSCAN, with high precision but low recall, allows more spam messages to evade detection and pass through the filters. In contrast, our approach with high precision and low recall demonstrates a conservative filtering strategy that accurately identifies legitimate messages while allowing some spam to bypass detection.

TABLE I
PRECISION, RECALL, BALANCED ACCURACY (ACC), SPD (FREE, WIN), EOD (FREE, WIN), AND TED (FREE, WIN) FOR OUR APPROACH (CEC), DBSCAN AND K-MEANS

Model	P	R	ACC	SPD	EOD	TED
CEC	1.00	0.68	0.85	0.13	0.02	0.00
DBSCAN	0.99	0.59	0.79	0.14	0.07	0.02
K-Means	0.82	0.44	0.71	0.14	0.06	0.08

Summary for RQ1: As highlighted in yellow, CEC surpassed DBSCAN and K-Means, achieving a balanced accuracy of 85% and a TED score of 0.00.

2) RQ2:: Can a context-driven clustering technique effectively automate the selection of representative samples to generate prompts to fine-tune large-language models in SMS spam detection?

Today, most applications rely on manual prompts to bridge the gap between human and LLM language to achieve the best performance. To address this problem in SMS spam detection, we used our CEC approach to help automate the selection of samples, which can be used as prompts for fine-tuning large-language models (LLMs). This eliminates the need for manual labeling and prompt generation, which is time-consuming and prone to human-induced bias. In this study, we used our CEC approach to select samples, which are used as input to fine-tune ChatGPT-4. CEC, together with ChatGPT-4, achieved a balanced accuracy of 98%, recall of 97%, and precision of 88%, showing the ability of ChatGPT-4 to learn from unbiased, high-quality examples from our CEC approach while maintaining fairness with an EOD and a TED of 0, as shown in Table II.

TABLE II
PRECISION, RECALL, BALANCED ACCURACY (ACC), SPD (FREE, WIN), EOD (FREE, WIN), AND TED (FREE, WIN) FOR CHATGPT-4 USING CEC FOR PROMPT SELECTION

Model	P	R	ACC	SPD	EOD	TED
ChatGPT-4	0.88	0.97	0.98	0.06	0.00	0.00

Summary for RQ2: As highlighted in yellow, when CEC is used to select samples as input to fine-tune ChatGPT-4, it achieves a higher balanced accuracy of 98%, an EOD score of 0.00, and a TED score of 0.00.

V. DISCUSSION

Our experimental results highlight the significant potential of CEC to effectively mitigate human-induced label bias by incorporating context-aware clustering and adaptive thresholding in SMS spam datasets. Compared to traditional methods like DBSCAN and K-means, which have lower recall and balanced accuracy, CEC demonstrates superior balanced accuracy, recall, treatment equality difference (TED), and equal opportunity difference (EOD), and also ensures balanced representation of spam-related context, reducing over-reliance on specific keywords. Using CEC to select samples for fine-tuning LLMs such as ChatGPT-4 further enhances fairness and classification accuracy, addressing the limitations of existing clustering methods. Our approach minimizes the dependence on human-induced labeled training data, providing a more adaptive and effi-

cient solution for rapidly evolving SMS spam patterns while surpassing traditional machine learning models.

VI. Threats to validity

While our Context-Enhanced Clustering (CEC) approach effectively mitigates human-induced label bias in SMS spam detection, certain factors may impact its generalizability and validity:

- The reliance on English-only SMS messages restricts the applicability of this approach to non-English languages. Spam characteristics, contextual meanings, and linguistic structures vary across different languages, potentially affecting model performance when applied to multilingual datasets.
- CEC relies on adaptive clustering with contextual weighting, which improves fairness but may be sensitive to hyperparameter selection (clustering thresholds, similarity metrics). Variability in parameter tuning could impact model outcomes, requiring further optimization strategies for different datasets and spam trends.
- SPD, EOD, and TE assess spam detection's fairness, minimizing classification disparities across different message types. However, these metrics alone may not capture all forms of subtle bias in spam detection.

VII. Conclusion and Future Work

Our research presents a novel approach to mitigate human-induced label bias in SMS spam detection using a context-enhanced clustering (CEC) framework. The study demonstrates that CEC and using CEC to select samples for fine-tuning large language models such as ChatGPT-4 achieved highly balanced accuracy, recall, and precision while maintaining low equal opportunity and treatment equality differences compared to the work of G.S. fuhnwi et al. [19]. This highlights the potential of CEC for efficient and accurate SMS spam detection, eliminating the need for traditional state-of-the-art machine learning or deep learning approaches that rely on large, labeled datasets, which are costly in terms of human labeling.

Future work will focus on extending the CEC approach to multilingual datasets, sentiment analysis, explore additional fairness evaluation frameworks and implementing real-time spam filtering capabilities.

References

[1] M. Min, J. J. Lee, and K. Lee, "Detecting illegal online gambling (IOG) services in the mobile environment," Security and Communication Networks, vol. 2022, no. 1, pp. 3286623, 2022.

[2] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise," Computer Networks, vol. 59, pp. 101–121, 20214, Elsevier.

[3] U. Mmaduekwe, "Bias and Fairness Issues in Artificial Intelligence-driven Cybersecurity," Current Journal of Applied Science and Technology, vol. 43, no. 6, pp.109–119, 2024.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A survey on bias and fairness in machine learning," ACM computing surveys (CSUR), vol. 54, no. 6, pp. 1–35, 2021, ACM New York, NY, USA.

[5] J.A. Hartigan and M.A. Wong "A k-means clustering algorithm," Applied statistics, vol. 28, no. 1, pp.100–108, 1979, USA.

[6] Fuhnwi, Gerard Shu and Agbaje, Janet O and Oshinubi, Kayode and Peter, Olumuyiwa James: An Empirical Study on Anomaly Detection Using Density-based and Representative-based Clustering Algorithms, Journal of the Nigerian Society of Physical Sciences, pp.1364-1364, (2023).

[7] J. M. Rao and D. H Reiley, "The economics of spam," Journal of Economic Perspectives, vol. 26, no. 3, pp. 87–110, 2012, American Economic Association.

[8] M. Labonne and S. Moran, "Spam-t5: Benchmarking large language models for few-shot email spam detection," arXiv preprint arXiv:2304.01238, 2023.

[9] N. K. Nagwani and A. Sharaff, "SMS spam filtering and thread identification using bi-level text classification and clustering techniques," Journal of Information Science, vol. 43, no. 1, pp. 75–87, 2017, SAGE Publications Sage UK: London, England.

[10] A. Shinde, E. Q. Shahra, S. Basurra, F. Saeed, A. A. AlSewari and W.A. Jabbar, "SMS Scam Detection Application Based on Optical Character Recognition for Image Data Using Unsupervised and Deep Semi-Supervised Learning," Sensor, vol. 24, no. 18, pp. 6084, 2024, MDPI.

[11] H. Baaqeel and R. Zagrouba, "Hybrid SMS spam filtering system using machine learning techniques," 2020 21st International Arab Conference on Information Technology (ACIT), pp. 1–8, 2020, IEEE.

[12] D. Pandya, "Spam detection using clustering-based SVM," Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence, pp. 12–18, 2019.

[13] H.P. Ahmad and S. Dang, "Performance Evaluation of Clustering Algorithm Using different dataset," International Journal of Advance Research in Computer Science and Management Studies, vol. 8, 2015.

[14] K. Fort, G. Adda and K. B. Cohen,"Amazon Mechanical Turk: Gold mine or coal mine?," Computational Linguistics, vol. 37, no. 2, pp. 413–420, 2011.

[15] S. Sun, K. Lei, Z. Xu, W. Jing and G. Sun, "Analysis of K-means and K-DBSCAN commonly used in data mining,"2023 International Conference on Intelligent Media, Big Data and Knowledge Mining (IMBDKM), pp. 37–41, 2023, IEEE.

[16] T. Almeida and J. Hidalgo, "SMS Spam Collection," UCI Machine Learning Repository, 2011. [Online]. Available: https://doi.org/10.24432/C5CC84.

[17] B. Steenhoek, M. M. Rahman, M. K. Roy, M. S. Alam, E. T. Barr and W. Le, "A Comprehensive Study of the Capabilities of Large Language Models for Vulnerability Detection," arXiv preprint arXiv:2403.17218, 2024.

[18] S. Verma and J. Rubin, "Fairness definitions explained,"Proceedings of the international workshop on software fairness, pp. 1–7, 2018.

[19] G. S. Fuhnwi, M. Revelle, B. Whitaker and C. Izurieta, "Using Large Language Models to Mitigate Human-Induced Bias in SMS Spam: An Empirical Approach," 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA, 2025, pp. 1-7, doi: 10.1109/ICAIC63015.2025.10848636.