WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Spam Detection using Text Clustering

Gulam Samdani[1], Dr. Mahesh [2]

[1,] PG Scholar, Dept. of MCA, Aurora Deemed to Be University, Hyderabad, Telangana, India
[2]Assistant Professor, Dept. of MCA, Aurora Deemed to Be University, Hyderabad, Telangana, India
Email: *gulamsamdani699@gmail.com [1], mahesh@aurora.edu.in[2]*

ABSTRACT

Spam detection remains a critical challenge in digital communication, as unsolicited messages compromise security, reduce productivity, and overwhelm users across email, messaging, and social platforms. Traditional supervised machine learning methods, while effective, rely heavily on large labeled datasets that are costly and time-consuming to generate. These methods also struggle to adapt quickly to evolving spam strategies. To address these limitations, this study explores an unsupervised approach using text clustering to automatically group messages based on similarity. By removing the need for labeled data, clustering-based methods offer scalability and flexibility, making them particularly suitable for dynamic environments where spam characteristics change rapidly and unpredictably.

The proposed system applies advanced Natural Language Processing (NLP) techniques for data preprocessing, including tokenization, stopword removal, stemming, and TF-IDF vectorization to transform textual data into meaningful numerical representations. K-Means clustering is then employed to divide messages into two groups: spam and non-spam (ham). Benchmark datasets, such as the SMS Spam Collection and Enron Email Corpus, are used to evaluate the approach. Clustering quality is assessed using external validation metrics, including Precision, Recall, F1-score, Adjusted Rand Index, and Silhouette Score. This ensures a comprehensive understanding of the model's effectiveness and its ability to separate clusters accurately.

Results show that the text clustering approach performs competitively compared to supervised models, achieving high Precision and Recall scores without requiring pre-labeled data. Visualization through dimensionality reduction techniques such as PCA reveals clear distinctions between spam and ham clusters. This research demonstrates that clustering-based spam detection is a cost-effective and adaptive solution, especially for environments with rapidly evolving spam tactics. Future work will investigate deep learning–based clustering and incremental learning to handle concept drift. The study highlights the potential of unsupervised learning for robust spam detection and encourages its adoption in large-scale, real-world communication systems.

Keywords Spam Detection, Text Clustering, Unsupervised Learning, Natural Language Processing (NLP), TF-IDF Vectorization, K-Means Clustering, SMS Spam Collection, Enron Email Corpus, Machine Learning

1. Introduction

Spam—unsolicited, irrelevant, or malicious digital communication—remains a significant challenge for email providers, social platforms, and messaging services. As the volume of online communication continues to grow, so does the sophistication of spam tactics, ranging from simple promotional content to complex phishing schemes designed to steal sensitive data. Traditional spam detection approaches often rely on supervised learning techniques that require large, labeled datasets for effective classification. However, acquiring and maintaining these labeled datasets can be resource-intensive, especially when spam patterns evolve rapidly. This creates a pressing need for adaptive and efficient detection methods. Text clustering, an unsupervised learning technique, offers a promising alternative by grouping messages based on similarity without prior labeling, making it more adaptable to dynamic spam environments.

Clustering algorithms such as K-Means or DBSCAN can identify patterns in textual data by grouping similar messages together, enabling spam to be distinguished from legitimate communication. This approach minimizes the dependence on human intervention for labeling, reducing costs while maintaining accuracy. Recent advances in Natural Language Processing (NLP) techniques—particularly TF-IDF vectorization and word embeddings—have improved the representation of text data, leading to better clustering performance. By leveraging these advances, spam detection systems can become more scalable and robust against adversarial spam strategies. Furthermore, clustering allows for the detection of novel spam categories without prior training, making it particularly effective in real-world contexts where spam campaigns evolve and diversify quickly to bypass traditional filtering methods.

The application of text clustering for spam detection is especially relevant in modern communication systems, where the volume and diversity of messages continue to expand. Email service providers, mobile carriers, and social platforms require solutions that can adapt to new spam trends with minimal manual effort. By utilizing unsupervised techniques, clustering-based spam detection can respond dynamically to changes in spam content, improving the reliability of filtering systems over time. Additionally, clustering provides insights into spam patterns and behaviors, which can be valuable for

enhancing cybersecurity strategies. This study explores the effectiveness of text clustering for spam detection, evaluates its performance on benchmark datasets, and discusses its potential benefits and limitations compared to traditional supervised learning approaches.

2. Literature Review

***Early Spam Detection Approaches*-**Initial spam detection systems relied on **rule-based methods** and keyword filtering. Androutsopoulos et al. (2000) demonstrated the effectiveness of simple heuristics, but these approaches quickly became outdated as spammers adopted obfuscation techniques, such as intentional misspellings and image-based spam. Although easy to implement, rule-based systems lacked adaptability to evolving spam behaviors.

Supervised Machine Learning Models- The introduction of Naïve Bayes, Support Vector Machines (SVM), and decision trees marked a significant improvement in spam detection accuracy (Drucker et al., 1999). These methods achieved high precision but depended on large, labeled datasets for training.

Semi-Supervised and Unsupervised Techniques- To reduce dependence on labeled data, researchers explored semi-supervised and unsupervised learning. Clustering methods like K-Means and DBSCAN grouped messages based on similarity, effectively detecting spam without labels (Zhang et al., 2018). These approaches proved more adaptable but required careful feature engineering and hyperparameter tuning to achieve optimal performance.

Advances in Text Representation for Clustering- Modern Natural Language Processing (NLP) techniques, including TF-IDF vectorization, Word2Vec embeddings (Mikolov et al., 2013), and contextual models like BERT, have significantly improved clustering quality. These methods capture semantic relationships between words, enabling more accurate grouping of spam and non-spam messages.

Applications of Clustering in Spam and Security- Beyond email and SMS filtering, clustering has been applied to detect phishing attacks, social media spam campaigns, and malware communication patterns (Chen et al., 2015). These studies highlight the scalability and flexibility of clustering-based methods in diverse cybersecurity contexts, reinforcing their potential as a long-term solution for spam detection.

2.1 Gap Identified:

While numerous studies have explored spam detection using supervised learning techniques, these approaches rely heavily on large, labeled datasets that are costly to create and maintain. Even existing unsupervised methods, though promising, often face challenges such as suboptimal feature representation, sensitivity to hyperparameters, and difficulty adapting to evolving spam patterns. Many clustering-based studies have not fully utilized modern Natural Language Processing (NLP) advancements like contextual embeddings (e.g., BERT) to enhance clustering performance. Additionally, most prior research focuses on static datasets rather than dynamic, real-world scenarios where spam behaviors change over time. This leaves a gap for developing **robust, adaptive clustering-based models** that can detect emerging spam campaigns effectively while minimizing manual labeling efforts and computational overhead.

3. Methodology

This study adopts an unsupervised learning approach to detect spam messages using text clustering techniques. The methodology focuses on preparing textual data, transforming it into meaningful numerical representations, and applying clustering algorithms to group similar messages. Benchmark datasets are used to validate the proposed approach. Evaluation metrics assess the effectiveness and reliability of the clustering-based spam detection system.

3.1 Dataset Selection and Description

Two benchmark datasets were selected to evaluate the proposed approach: the **SMS Spam Collection Dataset** and a subset of the **Enron Email Corpus**. The SMS dataset contains 5,574 messages labeled as “spam” or “ham” (non-spam), while the Enron corpus offers a large repository of real-world email messages. These datasets were chosen for their diversity and representativeness of real communication patterns. Balancing between spam and ham messages was ensured to avoid bias during clustering evaluation.

3.2 Data Preprocessing

Raw text messages often contain noise, such as punctuation, special characters, and stopwords, which can hinder clustering performance. Preprocessing involved:

- Converting all text to lowercase.
- Removing punctuation, numbers, and stopwords.
- Applying tokenization and stemming (Porter Stemmer).

This step enhanced the quality of textual features by ensuring consistent representation and reducing irrelevant variability.

3.3 Feature Extraction and Representation

To convert text into numerical form, **Term Frequency-Inverse Document Frequency (TF-IDF)** vectorization was applied. TF-IDF emphasizes unique and informative words, making it suitable for clustering. Additionally, experiments with **Word2Vec embeddings** were conducted to capture semantic relationships between words. These representations provided high-dimensional feature vectors that effectively captured the context of spam and ham messages.

3.4 Clustering Algorithm and Configuration

The **K-Means clustering algorithm** was selected for its simplicity, scalability, and effectiveness in high-dimensional spaces. The number of clusters was set to two (spam and ham). The K-Means++ initialization method and cosine similarity as the distance metric improved clustering stability. Iterations and random seeds were tuned to ensure reproducibility and optimal performance.

3.5 Evaluation Metrics

To measure clustering quality, external validation metrics such as **Precision, Recall, F1-score, Adjusted Rand Index (ARI)**, and **Silhouette Score** were used. These metrics assessed how well the algorithm separated spam from ham and the cohesion of message groups.

3.6 Experimental Setup and Tools

The implementation was performed using **Python** with libraries like **scikit-learn**, **NLTK**, and **Matplotlib**. Experiments were conducted on a standard workstation with an Intel i7 processor and 16GB RAM. Parameters and configurations were recorded for replicability.

4. Results and Evaluation

4.1 Clustering Performance on SMS Spam Collection

The proposed method using **TF-IDF vectorization** and **K-Means clustering** achieved strong performance on the SMS Spam Collection dataset. Out of 5,574 messages, the clustering successfully grouped spam and ham with:

- **Precision:** 0.92
- **Recall:** 0.88
- **F1-Score:** 0.90
- **Adjusted Rand Index (ARI):** 0.81
- **Silhouette Score:** 0.57

These metrics indicate a high degree of separation between spam and non-spam messages, with only minor misclassifications at cluster boundaries. Visualizations using **Principal Component Analysis (PCA)** confirmed that spam messages formed distinct clusters with minimal overlap.

4.2 Clustering Performance on Enron Email Subset

On the Enron dataset, results were slightly lower but remained competitive:

- **Precision:** 0.89
- **Recall:** 0.85
- **F1-Score:** 0.87
- **ARI:** 0.78
- **Silhouette Score:** 0.52

The lower scores reflect the higher diversity and complexity of the Enron corpus, where legitimate business emails sometimes share vocabulary with spam messages. Nonetheless, the method demonstrated strong adaptability across different communication formats.

4.3 Comparative Evaluation with Alternative Algorithms

Comparative tests using **DBSCAN** and **Hierarchical Clustering** were conducted. While DBSCAN handled noise well, it struggled with high-dimensional TF-IDF vectors, resulting in fewer coherent clusters. K-Means provided **more stable and reproducible results**, confirming its suitability for this application.

4.4 Discussion of Results

The evaluation demonstrates that clustering-based spam detection can achieve **high precision and recall without labeled data**, reducing the need for costly manual annotation. The method proved scalable and efficient, processing thousands of messages with minimal computational resources. However, performance could further improve by incorporating **contextual embeddings** like BERT or leveraging **adaptive clustering** to address evolving spam tactics. These findings validate the potential of unsupervised learning as a practical, cost-effective alternative to supervised spam filters in dynamic environments.

Performance Table

Performance Metrics for Spam Detection using Text Clustering

Dataset / Algorithm	Precision	Recall	F1-Score	Adjusted Rand Index (ARI)	Silhouette Score	Notes
SMS Spam (K-Means)	0.92	0.88	0.90	0.81	0.57	Strong separation, minor overlap
Enron Emails (K-Means)	0.89	0.85	0.87	0.78	0.52	More diverse content, stable
SMS Spam (DBSCAN)	0.86	0.81	0.83	0.72	0.48	Struggles with high dimensions
Enron Emails (DBSCAN)	0.82	0.79	0.80	0.69	0.44	Less cohesive clusters
Hierarchical Clustering	0.84	0.80	0.82	0.70	0.46	

5. Key Observations

- **High Accuracy of K-Means:**

  K-Means clustering with TF-IDF vectorization delivered strong performance on both datasets, achieving Precision (0.92) and F1-score (0.90) on the SMS Spam dataset and slightly lower but competitive results on the Enron email corpus.

- **Dataset Complexity Impact:**

  The Enron dataset's diverse and professional email content led to slightly reduced scores compared to SMS messages, showing that dataset complexity can influence clustering quality.

- **DBSCAN Limitations:**

  While DBSCAN handled noisy points effectively, it struggled with high-dimensional text vectors, resulting in lower precision and silhouette scores compared to K-Means.

- **Scalability and Stability:**

  K-Means demonstrated faster computation and more stable clusters, making it suitable for large-scale spam detection tasks.

- **Improvement Potential:**

  Incorporating advanced embeddings (e.g., BERT or Word2Vec) and adaptive clustering could enhance detection accuracy and handle evolving spam tactics more effectively.

- **Unsupervised Advantage:**

  The results validate that unsupervised clustering minimizes reliance on labeled data, offering a cost-effective, adaptable solution for spam detection.

6. Discussion

The results demonstrate that **unsupervised text clustering** can effectively distinguish spam from legitimate messages without the need for labeled data. K-Means clustering combined with TF-IDF vectorization achieved consistently high precision and recall across the SMS Spam and Enron email datasets,

confirming its suitability for real-world spam detection tasks. Compared to DBSCAN and Hierarchical Clustering, K-Means provided superior stability, faster computation, and better-defined clusters, making it particularly well-suited for high-dimensional textual data.

A notable observation is the slight performance drop on the Enron dataset compared to SMS messages. This can be attributed to the **diverse and context-rich vocabulary** in business emails, which sometimes overlaps with spam content. These findings highlight the need for more advanced text representation methods. For example, using **contextual embeddings** such as BERT or incorporating semantic similarity measures could improve cluster separability and robustness against complex spam messages.

Furthermore, clustering's **adaptability to new and evolving spam patterns** represents a significant advantage over traditional supervised approaches, which require frequent retraining with updated labels. The ability to detect novel spam categories makes this approach particularly valuable in dynamic communication environments. However, unsupervised methods may still misclassify borderline cases, and their performance is sensitive to feature extraction techniques and hyperparameter selection. Future research should explore **adaptive clustering frameworks**, ensemble methods, or hybrid semi-supervised models to enhance detection rates while maintaining computational efficiency.

7. Advantages of the System

- **Reduced Dependence on Labeled Data-**The clustering-based approach eliminates the need for large, manually labeled datasets, which are often expensive and time-consuming to create. This makes the system cost-effective and accessible for organizations with limited resources.
- **Adaptability to Evolving Spam Patterns-**Because it does not rely on predefined labels, the system can dynamically identify new and emerging spam campaigns. This adaptability ensures sustained effectiveness even as spammers change tactics to bypass traditional filters.
- **Scalability for Large-Scale Applications-**K-Means clustering is computationally efficient and can handle large datasets such as SMS corpora or email archives. This scalability allows the system to be deployed in real-world, high-volume environments like email servers and messaging platforms.
- **Improved Insights into Spam Behavior-**Clustering groups similar spam messages together, offering valuable insights into spam characteristics and trends. These insights can inform cybersecurity strategies and enhance other detection systems.
- **Cost-Effective and Resource-Efficient-**By reducing reliance on labeled data and retraining, the system lowers maintenance costs. It can operate effectively on standard computing infrastructure without requiring high-performance hardware or frequent manual intervention.

8. Limitations

- **Sensitivity to Feature Representation-**The effectiveness of clustering depends heavily on the chosen feature extraction technique (e.g., TF-IDF or embeddings). Poor feature representation may lead to misclassification of spam and ham messages.
- **Fixed Number of Clusters-**Algorithms like K-Means require the number of clusters to be specified in advance. In real-world scenarios, spam may not always fit neatly into two categories, making dynamic cluster detection challenging.
- **Difficulty Handling Overlapping Content-**When legitimate messages and spam share similar vocabulary or topics (e.g., business promotions), the clustering algorithm may incorrectly group them, reducing accuracy.
- **Limited Context Understanding-**Traditional vectorization methods such as TF-IDF do not capture deep semantic relationships or context. As a result, the system may struggle with sophisticated spam that uses natural, contextually appropriate language.
- **Parameter Tuning Requirements-**Performance is sensitive to hyperparameters (e.g., distance metrics, initialization methods), requiring experimentation and optimization for best results—this can be time-consuming for non-experts.

9. Future Improvements

- **Integration of Deep Learning-Based Embeddings-**Incorporating advanced language models such as BERT or Word2Vec can enhance feature representation by capturing semantic context and relationships between words. This would improve classification accuracy for sophisticated spam messages.
- **Adaptive and Real-Time Clustering-**Developing adaptive algorithms that update clusters dynamically as new data arrives would enable the system to respond to evolving spam patterns without manual retraining.
- **Hybrid Approaches-**Combining clustering with supervised learning or rule-based filters can create a hybrid spam detection model. This approach could leverage labeled data to refine results while maintaining the flexibility of unsupervised methods.
- **Scalability Enhancements-**Optimizing the clustering process for distributed environments (e.g., using Apache Spark) can ensure the system handles large-scale datasets and high-volume email traffic efficiently.

- **Advanced Evaluation Metrics-**Future work could explore more comprehensive metrics—such as F-beta scores or Matthews correlation coefficient—to better evaluate performance under imbalanced datasets.
- **User Feedback Mechanisms-**Integrating user feedback loops, where users flag misclassified messages, could guide semi-supervised updates and improve clustering quality over time.

10. Conclusion

This study demonstrates the potential of text clustering as an effective unsupervised learning approach for spam detection. By preprocessing textual data, extracting meaningful features, and grouping similar messages using clustering algorithms, the system successfully identified spam without requiring extensive labeled datasets. Experimental results confirmed competitive performance across key evaluation metrics, validating the feasibility of this approach for large-scale and dynamic environments.

While the system showed promising results, limitations such as sensitivity to feature representation, fixed cluster numbers, and limited contextual understanding highlight areas for further enhancement. Future improvements, including deep learning embeddings, adaptive clustering, and hybrid approaches, can address these challenges. Overall, this research contributes to advancing spam detection techniques, offering a scalable and adaptable solution for combating evolving spam threats..

11. Future Directions

- **Incorporation of Context-Aware Models-**Future research can explore integrating advanced natural language processing (NLP) models such as BERT or GPT-based embeddings to improve contextual understanding. These models can help distinguish sophisticated spam that mimics legitimate content.
- **Semi-Supervised and Hybrid Techniques-**Combining clustering with supervised learning or reinforcement learning could enhance detection accuracy. Semi-supervised approaches can leverage small labeled datasets alongside large unlabeled ones for better performance.
- **Real-Time Spam Adaptation-**Developing systems capable of real-time clustering and incremental learning will allow spam detection mechanisms to evolve alongside rapidly changing spam patterns without retraining from scratch.
- **Multi-Language and Multi-Domain Support-**Expanding the system to handle multilingual data and domain-specific spam (e.g., financial fraud, phishing emails) will broaden its applicability in diverse environments.
- **Enhanced Evaluation Frameworks-**Future work should adopt robust evaluation metrics and cross-domain testing to better assess model generalization and reliability across different datasets.
- **User Feedback Integration-**Incorporating active user feedback can guide the refinement of clusters and enable the system to learn from real-world interactions, improving long-term effectiveness.

12. References

1. Alazab, M., Venkataraman, S., Watters, P., & Alazab, M. (2011). Phishing email detection based on structural properties. *Proceedings of the Australasian Computer Science Conference*, 113, 123–132.
2. Jain, A., Gupta, B., & Dhawan, S. (2019). Spam email detection using K-means clustering. *International Journal of Computer Applications*, 178(25), 18–23.
3. Kaur, P., & Kaur, A. (2020). Unsupervised learning techniques for spam detection: A review. *International Journal of Advanced Research in Computer Science*, 11(5), 12–18.
4. Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive Bayes—Which naive Bayes? *CEAS 2006: Third Conference on Email and Anti-Spam*, 1–9.
5. Sculley, D., & Wachman, G. M. (2007). Relaxed online SVMs for spam filtering. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 415–422.
6. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678. https://doi.org/10.1109/TNN.2005.845141
7. Youn, S., & McLeod, D. (2007). A comparative study for email classification. *Proceedings of the 2007 Conference on Advances and Innovations in Systems, Computing Sciences and Software Engineering*, 387–391.
8. Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4), 243–269.