Spam Detection Using Large Datasets with Multilingual Support

Anil Kumar Jatra[1] (Student)
Master of Technology in Artificial Intelligence and Machine Learning

Kusum Sharma[2] (Guide)
ORCID ID: 0009-0005-3220-159X
Department of Computer Science and Technology
RSR Rungta College of Engineering & Technology Kohka Kurud Road Bhilai
Chhattisgarh, India

Abstract:- Spam detection in the era of big data requires scalable and efficient techniques, particularly when dealing with large datasets containing diverse languages. Traditional methods struggle to address the multilingual nature of spam, as language-specific approaches may not generalize well across different languages. This paper explores the establishment of a spam block method that leverages large, diverse datasets encompassing multiple languages. We employ advanced machine-learning techniques to handle the complexities of linguistic variations. By incorporating cross-lingual embeddings, transfer learning, and ensemble models, our system aims to detect spam content across various languages accurately. We highlight the importance of feature extraction, text preprocessing, and model adaptation in achieving robust multilingual spam detection. The proposed approach demonstrates improved performance in detecting spam messages while maintaining scalability and adaptability to new languages, providing a foundational framework for combating spam globally.

Keywords:- *Spam Detection, Multilingual Spam, Machine Learning, Cross-Lingual Embeddings, Transfer Learning, Ensemble Methods, Feature Extraction, Text Preprocessing, Model Adaptation, Large Datasets, and Language-Independent Spam Detection.*

I. INTRODUCTION

Spam detection is a crucial task, especially with the vast amount of data generated every day, including messages from emails, social media, and messaging apps. Traditional methods struggle to detect spam effectively when dealing with multiple languages. Most existing spam detection systems are built for specific languages, which means they don't work well for others, reducing their accuracy in identifying spam in diverse datasets. With communication platforms reaching a global audience, there's a growing need for systems that can detect spam in different languages, ensuring they are scalable and adaptable.

This research focuses on creating a spam detection system that can handle large datasets with multiple languages using machine learning techniques. By using methods like cross-lingual embeddings, transfer learning, and combining models, we aim to address the challenges posed by different languages. The goal is to build a system that is scalable and accurate in detecting spam across various languages. The paper covers designing, implementing, and testing such a system, focusing on tasks like extracting features, processing text, and adapting models to ensure effective multilingual spam detection.

II. LITERATURE REVIEW

This literature review summarizes various research works that investigates the utilization of machine learning techniques for finding SMS spam, highlighting their methods, datasets, results, and future scopes.

A. Studies on Machine Learning Models for SMS Spam Detection

➢ *SMS Spam Detection Using Naive Bayes and SVM*

- *Dataset*: SMS Spam Collection Dataset, Kaggle datasets.
- *Findings*: Naive Bayes and SVM demonstrate high efficiency in handling high-dimensional data, achieving accurate spam classification.
- *Future Scope*: Real-time deployment using lightweight frameworks like Flask.

➢ *Performance of Random Forest and SVM*

- *Reference*: Journal of Physics: Conference Series.
- *Findings*: Random Forest and SVM performed robustly, achieving up to 95% accuracy. Preprocessing techniques like TF-IDF significantly improved classification results.
- *Future Scope*: Expanding datasets and developing large-scale, standardized benchmarks.
- *Future Scope*: Expansion to multilingual datasets and exploration of deep learning methods.

➢ *Relevance Vector Machine (RVM)*

- *Reference*: Journal of Computational Analysis and Applications.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14609293

- *Findings*: RVM outperformed other models, achieving an F1 score of 97.6%. However, it required longer training times.
- *Future Scope*: Dataset expansion and advanced ensemble methods for real-time applications.

➢ *Hybrid Approaches Using Ensemble Techniques*

- *Reference*: IJNRD, Volume 9.
- *Findings*: KNN with Manhattan distance and Random Forest achieved a 97.78% accuracy.
- *Future Scope*: Integration of neural networks and exploration of fuzzy logic.

➢ *Optimizing SMS Spam Detection with Ensemble Learning*

- *Reference*: Journal of Computer Networks.
- *Findings*: SVM achieved the highest accuracy (98.57%) among classifiers. Ensemble methods enhanced prediction reliability.
- *Future Scope*: Addressing class imbalance with advanced techniques like SMOTE and expanding datasets for multilingual support.

B. *Advanced Techniques and Neural Network Applications*

➢ *Transformer-Based Embeddings*

- *Reference*: Sensors 2023.
- *Findings*: Combining GPT-3 embeddings with an ensemble of classifiers obtained 99.91% accuracy.
- *Future Scope*: Applying the model to diverse datasets, including non-English languages.

➢ *Hybrid CNN-LSTM Model*

- *Reference*: Future Internet 2020.
- *Findings*: Achieved an accuracy of 98.37% in spam detection for English and Arabic SMS datasets.
- *Future Scope*: Enhancing framework functionalities for smishing and phishing detection.

➢ *Content-Based Neural Networks*

- *Reference*: IJE Transactions B: Applications.
- *Findings*: Averaged Neural Network achieved 98.8% accuracy with robust preprocessing methods, including feature engineering for URLs and emojis.
- *Future Scope*: Expanding datasets and developing large-scale, standardized benchmarks.

C. *Emerging Technologies and Future Directions*

➢ *Blockchain Integration for Spam Detection*

- *Reference*: IJISAE, 2024.
- *Findings*: Combining blockchain with machine learning ensures data transparency and integrity while maintaining high classification accuracy.
- *Future Scope*: Scaling blockchain integration for real-time systems and improving algorithm efficiency.

➢ *Bio-Inspired Algorithms*

- *Reference*: IEEE Access 2017.
- *Findings*: Techniques like Artificial Bee Colony and Cuckoo Search hold promise but remain underexplored for spam classification.
- *Future Scope*: Optimization of these algorithms and hybrid implementations.

➢ *Deep Learning for Multilingual Spam Detection*

- *Reference*: Hindawi Applied Computational Intelligence and Soft Computing.
- *Findings*: SVM outperformed CNN with 99.6% accuracy in SMS spam detection.
- *Future Scope*: Incorporating ensemble methods and expanding experiments to larger datasets.

D. *Evaluation Metrics and Dataset Limitations*

➢ *Common Datasets Used*

- UCI SMS Spam Collection, Kaggle, and other publicly available datasets dominated research efforts.
- Issues: Class imbalance (more ham messages than spam) and limited linguistic diversity.

➢ *Model Assessment using Metrics*

- Accuracy, Precision, Recall, and F1 Score were the most commonly used metrics.
- Challenge: Lack of standardized evaluation methods across studies.

E. *Summary and Recommendations*

While traditional algorithms like machine learning such as the Support vector machine, Random Forest method, and Naïve Bayes technique remain highly effective, advanced techniques like hybrid CNN-LSTM models and transformer-based embeddings have set new benchmarks in spam detection. Future work can be done by:

- Enlarging datasets to include diverse languages and formats.
- Exploring advanced learning models like bio-inspired methods and deep-learning techniques.
- Enhancing real-time deployment efficiency through lightweight and scalable frameworks.

This Related work addresses the challenges and growth in SMS spam detection, offering a foundation for further exploration.

III. PROBLEM IDENTIFICATION

➢ *Challenges in Model Adaptability:*

- Spam detection models require continuous updating to adapt to evolving spam techniques.
- Generalization across different languages, regions, and datasets is limited.

➢ *Handling Imbalanced Datasets:*

- Imbalanced datasets (more "ham" than "spam") pose challenges for model performance.

➢ *Preprocessing Variability:*

- Effective preprocessing (e.g., tokenization, stop-word removal, feature extraction) is essential but varies across studies, impacting model performance.

➢ *Feature Engineering Complexity:*

- Selection and optimization of feature parameters, such as message length and word frequency, significantly influence results.

➢ *Real-Time Deployment:*

- Many models lack real-time applicability due to computational or latency issues.

➢ *Limited Use of Advanced Techniques:*

- Limited exploration of advanced methods such as deep learning, hybrid approaches, and bio-inspired algorithms.

➢ *Evaluation Metrics and Consistency:*

- Lack of standardized evaluation metrics across studies, leading to challenges in comparing results.

IV. PROPOSED METHOD

To detect spam in multiple languages, we proposed the techniques that utilizes machine learning, especially ensemble classifiers, to improve accuracy and scalability.

➢ *Data Preprocessing & Gathering of Data:*

- **Gathering of data:** Gather large datasets that contain spam messages from different sources like emails, social media, and messaging apps.
- **Method of Preprocessing:** the data is cleaned by eleminating unnecessary information, and noise, and making sure the text is consistent across languages. This includes tokenization (breaking text into parts), stemming (reducing words to their roots), and normalization (making the text uniform).

➢ *Feature Extraction:*

- Extract features from the text, like word n-grams (combining words), character-level n-grams, and frequency-based features.

Methods like TF-IDF and word embeddings techniques is utilized in the presentation of texts that explains the meaning and works across different languages.

➢ *Model Development:*

- **Base Models:** Build models using simpler algorithms, such as Logistic Regression, for efficiency and Gradient-Boosting methods or the Random Forest technique to seize complicated patterns.
- **Ensemble Model:** Combine predictions from multiple root models approaching methods such as **voting** and **weighted average** to improve efficiency by combining different models.

➢ *Cross-Lingual Embeddings and Transfer Learning:*

- Use pre-trained multilingual word embeddings like **FastText** or **mBERT (multilingual BERT)** to understand relationships between words in different languages.
- Apply transfer learning by using models trained on one language to help understand other languages.

➢ *Ensemble Model Implementation:*

- Combine outputs from different base models using techniques like **Voting**, **Stacking**, or **Weighted Averaging** to improve accuracy.
- Use ensemble models like **Random Forest with stacking** or **XG Boost** with different base classifiers to enhance performance.

➢ *Evaluation of Models:*

- Performance metrics are used such as accuracy, precision, F1-score, and Recall across different languages.
- Cross-validation methods are implemented to check whether the model works well with different language data or not.

➢ *Adaptation and Scalability:*

- Continuously improve the system by adding more data and training on new languages.
- Ensure the system works efficiently with large datasets and different languages.

This method combines machine learning techniques, especially ensemble models, to detect spam more accurately and efficiently in multiple languages.

V. BACKGROUND WORK ON AN ENSEMBLE MODEL

To detect spam in multiple languages, we propose a method that uses machine learning techniques, especially ensemble models, to improve accuracy and scalability. In Ensemble models by combining strengths of many algorithms for improving comprehensive performance and robustness in SMS spam detection.

➢ *Data Collection*

- Collect messages labeled as spam or ham.
- Use universal datasets like:
- **SMS Spam Collection Dataset**: Popular dataset with labeled SMS messages.
- **Kaggle Datasets**: Platforms like Kaggle offer a variety of datasets for spam detection. Ensure data is accurate and properly labeled.

➢ *Method of Data Preprocessing*

- **Tokenization**: Breaking down the messages or texts into smaller pieces.
- **Cleaning of texts**: unnecessary texts or words, punctuation, and special characters are removed.
- **Handle Missing Data**: Ensure there are no gaps in the data; fix them if found.
- **Vectorization process**: the process of converting text into numbers using methods like TF-IDF (word embeddings).

➢ *Feature Selection method*

Selection of useful features to help in identifying spam messages, like:

- The length of Message.
- Existence of certain spam-related texts and patterns.
- Occurrence of specific phrases or words.

➢ *Selection of Models*

For text classification we select a machine learning Techniques like:

- Gradient Boosting Machines (GBM)
- Naïve Bayes Algorithm
- Decision Tree classification
- Random Forests method
- Support Vector Machines
- Ensemble Classifiers
- Logistic Regression method

➢ *Training of Models*

- We divide the datasets into two parts i.e. training and testing.
- Train the selected model using the training data.

➢ *Evaluation of Models*

- Checking the model on testing data.
- Checking performance metrics like accuracy, precision, recall, and F1-score.

➢ *Performing Hyperparameter Tunning*

- Adjusting some settings to our model to improve its efficiency.
- The cross-validation method is used to find the best parameters while avoiding overfitting.

➢ *Model Deployment*

- Implement a trained model to categorize messages as spam or ham in real-time.
- Allow users to see the classification results and provide feedback.

To understand the model's efficiency we can also include insight tools.

VI. MODEL DESIGN

➢ *Datasets*

In SMS spam detection, datasets play an important role in training the machine learning models. These datasets contain labeled messages contained as spam or ham. The machine learning models learn patterns and characteristics like word frequencies and text structures from this data to differentiate spam from non-spam messages. As new messages are analyzed, the model applies these patterns to predict if they are spam. Regularly updating the data enhances the model's accuracy and capability to adapt to new types of spam.

➢ *Data Extraction*

Data extraction involves collecting datasets of messages labeled as spam or ham. After cleaning the data we prepare it for analysis using methods like TF-IDF to extract features. Then we split the datasets into two parts (training and testing). The testing data evaluates how well the model performs, and modifications are done to enhance its efficiency. Once trained, the model is set up and categorizes new messages.

➢ *Data Visualization or Exploratory Data Analysis (EDA)*

EDA requires analyzing datasets visuals and statistics to understand its key properties. This step uses charts like histograms or scatter plots to identify patterns, trends, or anomalies in the data. It helps decide the best approaches.

➢ *Feature Engineering*

Feature engineering involves creating or selecting the most useful information from the raw data to improve model performance. This includes:

- Choosing relevant features (e.g., message length or specific keywords).
- Handling missing data.
- Converting text or categories into numerical values.
- Scaling features to ensure consistency. The goal is to provide the model with the most meaningful inputs for better predictions.

➢ *Model Building*

Building an ensemble classifier requies training many models and combined their predictions to obtained best performance as compared to other models. Ensemble techniques like bagging (i.e. Random Forest), boosting (i.e. Gradient Boosting), or stacking, leverage the performance of diverse model to increase in classification accuracy. These classifiers work by aggregating the predictions from multiple base learners, which may include decision trees, logistic regression, or other algorithms. The ensemble calculates the likelihood of a message being spam or non-spam by integrating predictions from these models, often using majority voting or weighted averages. This method is widely used for tasks like spam detection because it provides high accuracy and robustness by reducing over-fitting and leveraging diverse perspectives on the data.

➢ *Model Evaluation*

Evaluating the model involves measuring how well it classifies SMS messages. Metrics like:

- **Accuracy**: Overall correctness of the model.
- **Precision**: It shows the metrics that calculate how the model correctly gives predictions in positive terms.
- **Recall**: It is the metrics that calculate model positive instances from the datasets.
- **F1-Score**: These metrics help ensure the model reliably identifies spam while minimizing errors and also show the balance between recall and precision.

➢ *Predictions*

The trained model then analyzes new messages and divides them into spam or ham. It uses features and patterns learned during training to assign a label or probability to each message. This step is crucial for real-time spam detection, ensuring incoming messages are classified quickly to protect users from spam.

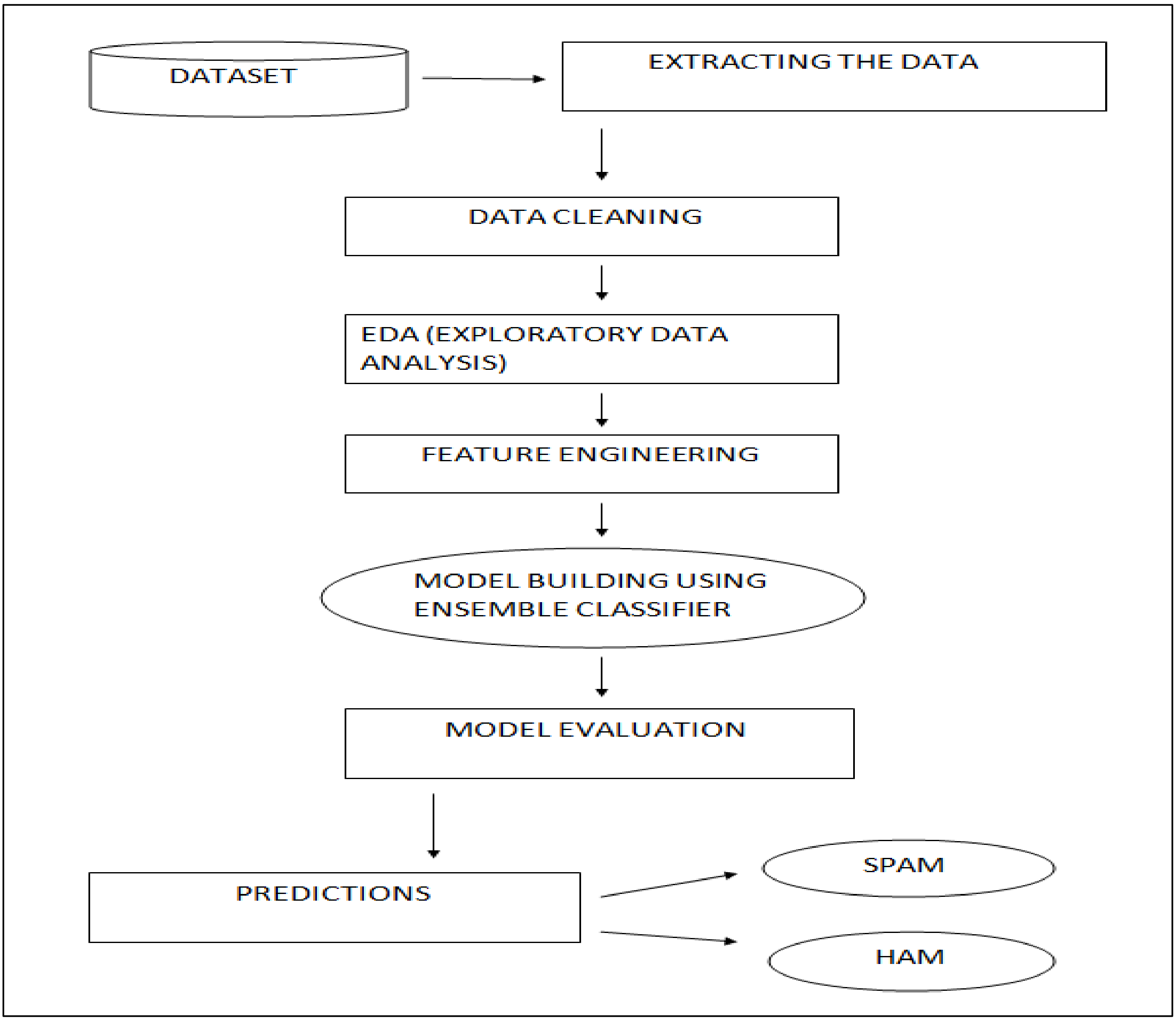

Fig 1 Ensemble Model Development

VII. CONCLUSION

This research tackles the challenge of detecting spam in multiple languages by creating a system that is accurate, scalable, and efficient. It uses advanced machine learning methods like Ensemble classifiers, cross-language embeddings, transfer learning, and combining multiple models to improve performance on large and diverse datasets.

➢ *Key Points Include:*

- Better Multilingual Support: The system works well with different languages, making spam detection more effective worldwide.
- Advanced Methods: Using modern techniques like deep learning and combining models, the system achieves high accuracy and reliability.
- Scalability and Real-Time Use: The system can handle large datasets and adapt to new languages and changing spam patterns quickly.
- Future Possibilities: The research suggests expanding datasets, using new technologies like blockchain, and exploring nature-inspired algorithms to improve spam detection further.

This study provides a strong foundation for global spam detection, ensuring it works accurately across languages and stays adaptable to new challenges.

REFERENCES

[1]. Shreya Menthe, Kanish Rawal, Mrudula Hirave, A.J.Patil, “SMS spam detection using machine learning” DOI: 10.17148/IJARCCE.2024. 13307

[2]. Suparna DasGupta, Soumyabrata Saha, Suman Kumar Das, “SMS spam detection using machine learning” Journal of Physics: Conference Series DOI: 10.1088/1742-6596/1797/1/012017

[3]. Ravi H Gedam, Sumit Kumar Banchhor,” Sms spam detection using machine learning” Journal of Computational Analysis and Applications Volume 33, No. 4, 2024

[4]. Arpita Laxman Gawade, Sneha Sagar Shinde, Samruddhi Gajanan Sawant, Rutuja Santosh Chougule, Mrs Almas Amol Mahaldar “A Research Paper of SMS Spam Detection” 2024 IJNRD, Volume 9, Issue 3-03-2024, ISSN: 2456-4184 | IJNRD.ORG

[5]. Harshit Kumar Simbal, Aaryan Sharma, Smriti Kumari, Gautam Kumar, Harshvardhan Kumar,” Spam Sms Classifier Using Machine Learning Algorithms” IJFMR240219483, Volume 6, Issue 2, March-April 2024

[6]. Gregorius Airlangga, “Optimizing SMS Spam Detection Using Machine Learning: A Comparative Analysis of Ensemble and Traditional Classifiers” Journal of Computer Networks, Architecture, and High-Performance Computing, Volume 6, Number 4, October 2024 DOI: 10.47709/cnahpc. v6i4.482

[7]. Shafi’l Muhammad Abdulhamid, (Member, IEEE), Muhammad Shafie Abd Latiff, Haruna Chiroma, (Member, IEEE), Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I. Abubakar, (Member, IEEE), and Tutut Herawan, “A Review on Mobile SMS Spam Filtering Techniques”IEEE Access Published: February 13, 2017

[8]. Pavas Navaney, Ajay Rana, Gaurav Dubey, “SMS Spam Filtering using Supervised Machine Learning Algorithms” Conference Paper DOI: 10.1109/CONFLUENCE. 2018.8442564

[9]. Pradeep K.B, “Sms spam detection using machine learning and deep learning techniques”, Published: May 2022

[10]. B Sai Deepthi, K Sudheer Kumar, CH B M Swaroop, K Satya Sudheer, “Sms spam filtering using machine learning” JETIR, May 2024, Volume 11, Issue 5 Sixth International Conference on Computing Methodologies and Communication (ICCMC 2022)

[11]. Mr. Ravi H. Gedam, Dr. Sumit Kumar Banchhor, “An Enhanced SMS Spam Detection Framework Using Blockchain and Machine Learning” IJISAE, 2024, Volume 12(22s), Pages 728–739

[12]. Samadhan Nagre, “Mobile SMS Spam Detection using Machine Learning Techniques” 2018 JETIR December 2018, Volume 5, Issue 12

[13]. Manas Ranjan Bishi, N Sardhak Manikanta, G Hari Surya Bharadwaj, P Siva Krishna Teja, Dr G Rama Koteswara Rao, “Optimizing SMS Spam Detection: Leveraging the Strength of a Voting Classifier Ensemble” IJISAE, 2024, Volume 12(3), Pages 2458–2469

[14]. Ahmed Alzahrani, “Explainable AI-based Framework for Efficient Detection of Spam from Text Using an Enhanced Ensemble Technique”, Engineering, Technology & Applied Science Research Volume 14, No. 4, 2024, Pages 15596-15601

[15]. Shushanta Pudasainia, Aman Shakyaa, ∗, Sanjeeb Prasad Pandeya, Prakriti Paudelb, Sunil Ghimirec, Prabhat Ale, “SMS Spam Detection using Relevance Vector Machine” 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023)

[16]. Abdallah Ghourabi, Manar Alohaly, “Enhancing Spam Message Classification and Detection Using Transformer-Based Embedding and Ensemble Learning” Sensors 2023, Volume 23, Article 3861DOI: 10.3390/s23083861

[17]. Abdallah Ghourabi, Mahmood A. Mahmood, Qusay M. Alzubi, “A Hybrid CNN-LSTM Model for SMS Spam Detection in Arabic and English Messages” Future Internet 2020, Volume 12, Article 156 DOI: 10.3390/fi12090156

[18]. Mr. E.Sankar, Y Y S Shekhar Babu, M.Tridev, “Sms spam detection using machine learning” International Journal of Scientific Research in Engineering and Management Volume 7, Issue 4, April 2023

[19]. Umair Maqsood, Saif Ur Rehman, Tariq Ali, Khalid Mahmood, Tahani Alsaedi, Mahwish Kundi, "An Intelligent Framework Based on Deep Learning for SMS and e-mail Spam Detection" Hindawi Applied Computational Intelligence and Soft Computing, Volume 2023 DOI: 10.1155/2023/6648970

[20]. Suvarna M, Sanjeev J R, Kiran K, Ganjendran, "Sms spam detection using machine learning" DOI: 10.17148/IARJSET.2024.11440

[21]. Nisha Wilvicta, Pradeep N, Tharun R, Mohammed Tousif, "Sms spam detection using machine learning" International Journal of Advances in Engineering Architecture Science and Technology DOI: 12.2023 13677758/IJAEAST. 2023.10.0001

[22]. Humaira Yasmin Aliza, Kazi Aahala Nagary, Eshtiak Ahmed, Kazi Mumtahina Puspita, Khadiza Akter Rimi, Ankit Khater, Fahad Faisal, "A Comparative Analysis of SMS Spam Detection Employing Machine Learning Methods" Proceedings of the

[23]. Andrew Kipkebut, Moses Thiga, Elizabeth Okumu, "Machine Learning Sms Spam Detection Model" Kabarak University International Conference on Computing and Information Systems, October 14–15, 2019

[24]. Samadhan M. Nagare, Pratibha P. Dapke, Syed Ahteshamuddin Quadri, Sagar B. Bandal, Manasi Ram Baheti, "A Review on Various Approaches on Spam Detection of Mobile Phone SMS" International Journal for Research in Engineering Applications & Management (IJREAM) ISSN: 2454-9150, Volume 9, Issue 2, May 2023

[25]. Luo GuangJun, Shah Nazir, Habib Ullah Khan, Amin Ul Haq, "Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms" Hindawi Security and Communication Networks, Volume 2020 DOI: 10.1155/2020/8873639