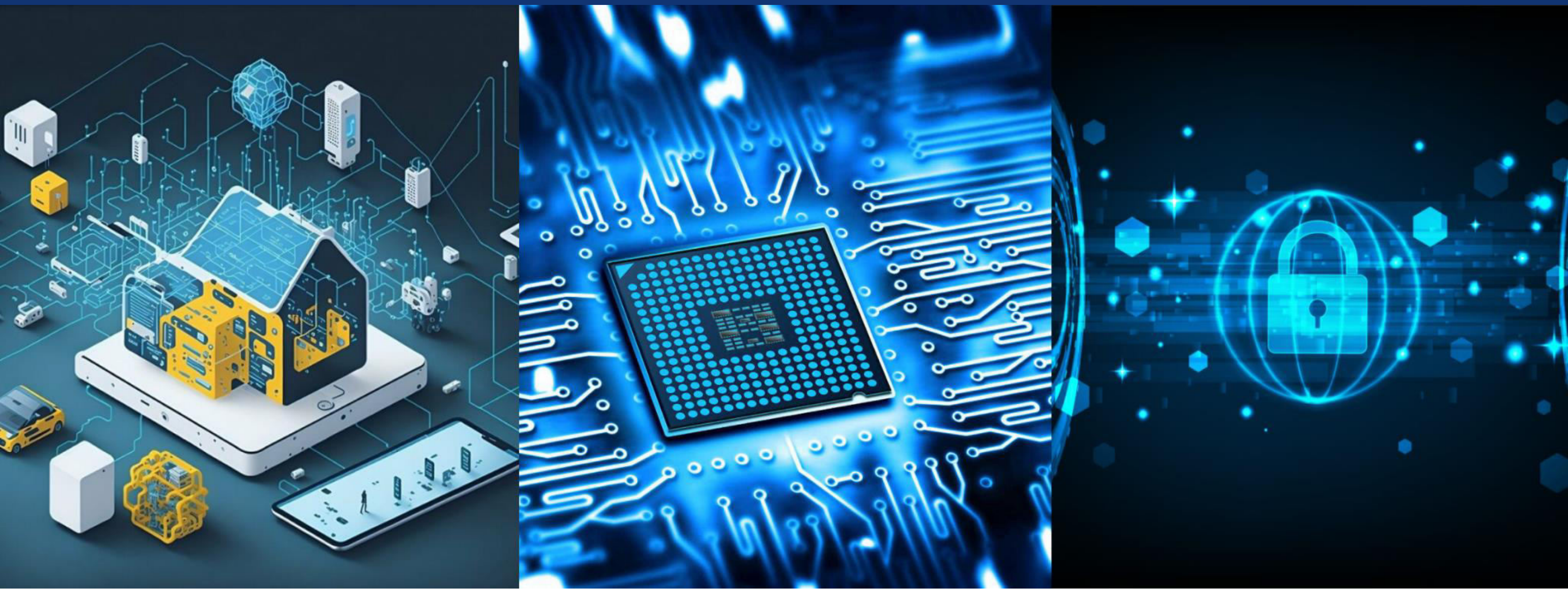
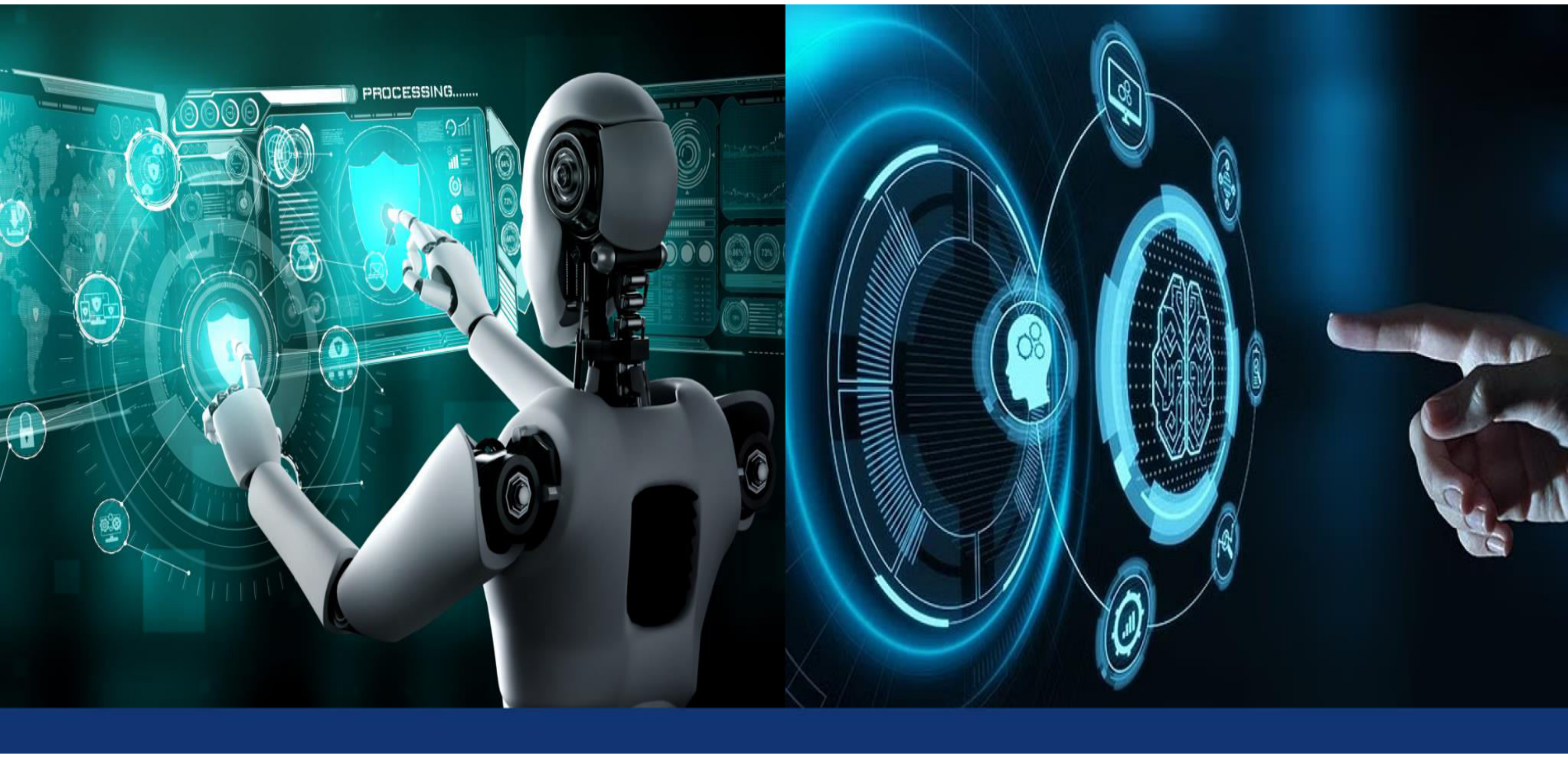
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 3, March 2025

www.ijircce.com ijircce@gmail.com +91-9940572462 +91 63819 07438

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Spam Detection on YouTube Video Comments Using Deep Learning Approaches

Sakshi Dodke[1], Kalyani Pachbhai, Dr. Pravin Game[3], Dr. Mubin Tamboli[4]

UG Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India [1,2]

Associate Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India [,3,4]

ABSTRACT: YouTube faces growing challenges with spam in comments, likes, and subscriptions, disrupting user experience and content integrity. Traditional rule-based spam detection methods are ineffective against evolving spam tactics. This study explores Convolutional Neural Networks (CNNs) for spam detection, leveraging their ability to process large-scale text and image data with high accuracy. The proposed CNN-based model achieves impressive accuracy rates up to 98.57%. Additionally, the research reviews various deep learning models, emphasizing adaptability, real-time implementation, and scalability for platforms like YouTube. The findings highlight CNNs' potential to enhance spam detection systems, ensuring more secure and reliable online environments.

KEYWORDS: YouTube, Spam Detection, Comments Spam, Deep Learning, Convolutional Neural Networks (CNNs), Spam Tactics Adaptation, Scalable Solutions, Precision and Resilience.

I. INTRODUCTION

In today's digital world, YouTube has become a huge platform where people share and watch videos every day. But with its popularity comes a big problem i.e. spam comments. These are unwanted messages that often include fake links, scams, or random promotions. They clutter up the comment sections, making it hard for real users to have meaningful conversations. They also hurt content creators by making their videos look less trustworthy. YouTube used to rely on simple rule-based systems to catch spam, like flagging certain words or limiting links. But spammers have gotten smarter, finding ways around these rules by changing spellings, using special characters, or creating fake accounts that act like real people. To solve this, more advanced tools are needed like machine learning and deep learning. These technologies can learn patterns from large amounts of data and improve over time. In particular, combining Convolutional Neural Networks (CNNs), which spot patterns in text, with Long Short-Term Memory (LSTM) networks, which understand the flow and meaning of language, makes for a powerful spam detection system. This study introduces a deep learning model that uses both CNN and LSTM to detect spam in YouTube comments, working quickly and accurately even as spammers change their tricks. The system is trained on a large variety of comments, so it can handle different languages and writing styles. It's built to keep learning and improving, making it a reliable solution for spotting spam as it evolves. The goal is to provide a scalable, efficient, and accurate solution that not only helps maintain the integrity of the platform but also supports content creators and viewers in having a better online experience. By testing the system in real-world situations, this study shows that it can handle large amounts of data while staying accurate and fast, helping YouTube stay trustworthy and user-friendly.

Specific Contributions

This paper makes the following contributions:

- Proposing a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for effective spam detection in YouTube comments.
- Implementing adaptive learning capabilities to enable the model to evolve alongside new spam strategies, maintaining high detection accuracy over time.
- Evaluating the performance of the proposed model using metrics such as accuracy, precision, recall, and F1-score

- Demonstrating the scalability and efficiency of the model for real-time spam detection on large platforms like YouTube, contributing to a safer and more trustworthy user experience.

By combining CNN and LSTM architectures, adaptive learning, and robust evaluation metrics, this paper aims to advance spam detection in YouTube comments and contribute to the development of more accurate and scalable content moderation tools.

II. LITERATURE SURVEY

In [1], machine learning techniques were applied to filter out spam in emails and IoT systems. These methods showed better accuracy and could handle large amounts of data. However, they also need constant updates and retraining to keep up with new spamming methods. In [9], the focus was on improving email spam detection by selecting the best features for machine learning models. This helped make the models more efficient, but there's a risk of accidentally removing important information during feature selection. Another approach in [10] combined deep learning with optimization techniques to improve spam detection. This method outperformed traditional techniques, although it required a lot more computing power.

In [3], the authors reviewed how machine learning can be used to spot fake reviews. Their study highlighted the difficulty in creating reliable datasets and accurately labeling data as spam or not. Despite these challenges, they found that using ensemble learning and good feature engineering helped improve the accuracy of detecting fake reviews and could be applied across different platforms.

Spam detection on social media, especially on Twitter, has also been a major focus. In [6], researchers applied machine learning models to detect spam in near real-time. These models were scalable and could handle a lot of data, but dealing with the fast pace and huge volume of tweets was still a big challenge. Sun et al. also worked on this issue by developing a framework that combined feature engineering with machine learning algorithms like logistic regression, Naive Bayes, random forests, CNNs, and LSTMs. Their system used tweet content, user account details, and behavior patterns to tell the difference between spam and real tweets. Still, they pointed out that keeping up with changing spam tactics and ensuring scalability remains tough.

Transformers are another promising approach. In [8], transformer models, which are typically used in computer vision, were discussed as a potential tool for spam detection in multimedia content. These models are flexible and powerful but require a lot of computing resources to train. Meanwhile, in [18], a survey looked at different spam detection methods on Twitter. These methods were grouped into content-based (looking at text and links), behavior-based (analyzing user actions), and graph-based approaches (examining network relationships), each offering unique advantages in spotting spam accounts.

In [20], a detailed comparison of deep learning models for spam detection in YouTube comments was carried out. Models like MLP, CNN, LSTM, BiLSTM, GRU, and Attention Mechanisms were evaluated. These models proved useful for detecting tricky spam messages and adapting to changing spam techniques, making them good choices for managing user comments on video platforms.

III. METHODOLOGY

In this, we propose an ensemble learning approach to improve the accuracy and reliability of spam detection on YouTube. Our method combines various machine learning algorithms, including Gaussian Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest, Decision Tree, and a Voting Classifier. By integrating these models, we aim to leverage their individual strengths and enhance the overall prediction accuracy. YouTube, as one of the largest video-sharing platforms, continues to struggle with spam comments that disrupt user interaction and can expose viewers to harmful or misleading content. Traditional rule-based spam filters are no longer sufficient as spammers constantly change their tactics. To address this issue, we incorporate deep learning techniques especially Convolutional Neural Networks (CNNs) and CNN-LSTM models to create a more adaptive and accurate spam detection system. Our system follows a streamlined process that includes data preprocessing, model development, and evaluation. We start by cleaning and

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

scaling the data, followed by Bayesian optimization to fine-tune model parameters. Cross-validation ensures robust evaluation, and the final ensemble model uses weighted voting to prioritize stronger classifiers. This approach not only increases accuracy but also helps the system adapt to evolving spam patterns, making it more reliable for real-world use.

The system design and architecture for the provided diagram consists of three major modules.

1. **Data Preprocessing:** The module starts with the YouTube Comments Dataset as input, where data cleaning removes noise and formatting issues, and tokenization converts the text into numerical form suitable for processing.
2. **Model Development:** Module consists of two parallel approaches. The first one uses a CNN model while the second combines CNN with an LSTM model. Both models are trained on the pre-processed data to extract the patterns and understand sequential dependencies in the comments.
3. **Evaluation Module:** evaluates trained models by accuracy, precision, and recall metrics. The best model selected was the one that has passed in the evaluation process. A module output from this was used by the Spam Detection Results Module to come up with a result of spam or not.

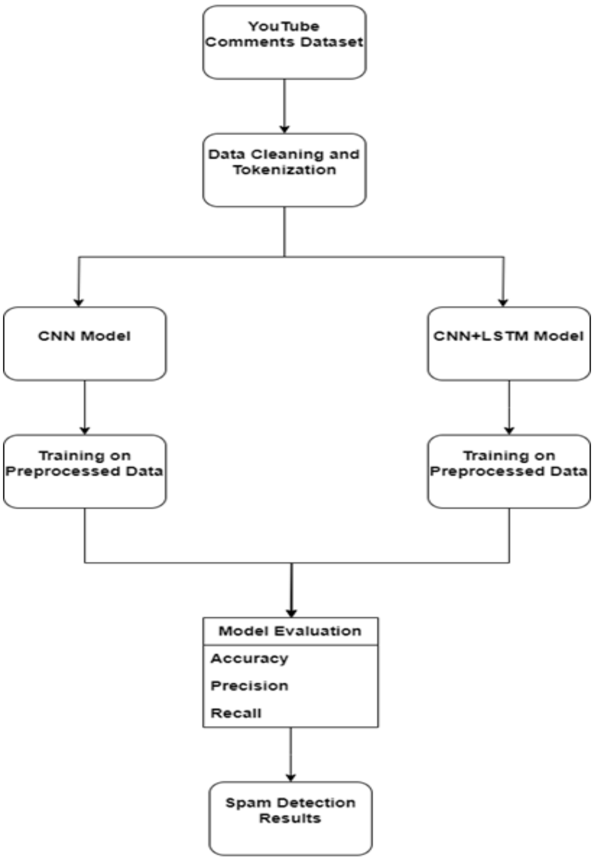


Fig. 1. System Architecture

A. Experimental Setup

- Dataset: The dataset consists of YouTube comments scraped from videos by popular artists such as Psy, Katy Perry, LMFAO, Eminem, and Shakira. These videos represent a variety of content, ensuring a diverse range of comment styles and topics. Each comment in the dataset is binary-labelled as either "spam" (1) or "ham" (0), where "spam" refers to irrelevant or malicious content and "ham" refers to irrelevant or malicious legitimate.

- Libraries: TensorFlow and Keras for building and training deep learning models, NumPy for numerical operations and data manipulation, Pandas for data preprocessing and handling structured data, Matplotlib and Seaborn for data visualization and performance evaluation and Scikit-learn for model evaluation and performance metrics
- Model Selection: Base models include Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory (LSTM) networks for sequential data analysis, configured to handle complex patterns and contextual information in spam comments.
- Stacking Classifier: In this case, the predictions of several deep learning models (such as LSTM, CNN, and Transformer-based models) are combined into a final model using a meta-learner.
- Model Evaluation: The performance of the proposed hybrid model is evaluated using accuracy, precision, recall, and F1-score metrics on the test dataset. Experimental results demonstrate the model's effectiveness in detecting spam comments while maintaining scalability for real-time application on large-scale platforms like YouTube.

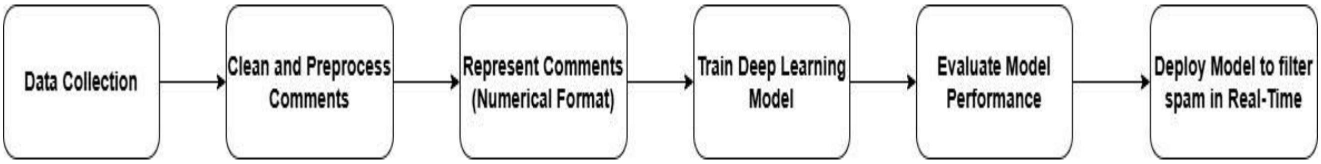

Data Collection
Clean and Preprocess Comments
Represent Comments (Numerical Format)
Train Deep Learning Model
Evaluate Model Performance
Deploy Model to filter spam in Real-Time

Fig. 2. Flowchart of Implementation

IV. RESULT

The Convolutional Neural Network (CNN) model achieved an impressive accuracy of 98.57% in detecting spam comments. With precision and recall scores of 97.22%, the model demonstrated high reliability in differentiating between spam and non-spam content. CNN's ability to capture complex patterns and contextual features makes it highly effective and efficient for spam detection tasks.

Table 1: Experimental results obtained using deep learning model for youtube02_katyperry.

Model	Accuracy	Precision	Recall	F1 Score
FNN	0.9429	0.9444	0.9444	0.9444
RNN	0.8286	0.7429	0.7222	0.7324
CNN	**0.9857**	**0.9722**	**0.9722**	**0.9722**
Bidirectional LSTM	0.9429	0.8974	0.9333	0.9429
CNN+LSTM	0.9429	0.6923	1.0000	0.8182

In Table 1, each individual model was evaluated for its performance. The results are as follows:

- Convolutional Neural Network (CNN) was the standout model, achieving the highest accuracy of 98.57%. Its precision and recall were both 97.22%, demonstrating its exceptional ability to accurately classify spam and non-spam comments.
- Feedforward Neural Network (FNN) achieved an accuracy of 94.29%, with precision, recall, and F1-score all at 94.44%. This indicates consistent and reliable performance in detecting spam content.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Bidirectional LSTM reached an accuracy of 94.29%, with a precision of 89.74% and recall of 93.33%, showing strong capabilities in understanding the sequential nature of spam comments.
- Recurrent Neural Network (RNN) had a lower accuracy of 82.86%, with a precision of 74.29% and recall of 72.22%, suggesting it struggled to effectively detect spam compared to other models.
- CNN + LSTM (Hybrid Model) also achieved an accuracy of 94.29%, with a perfect recall of 100%, but a lower precision of 69.23%, indicating it detected all spam comments but had a higher false positive rate.

Among these, the Convolutional Neural Network (CNN) proved to be the best model due to its ability to capture complex patterns and contextual relationships within the comment data. This capability contributed to its superior accuracy, making it particularly effective in differentiating between spam and non-spam comments.

Overall, the model's high accuracy highlights its potential as a valuable tool for spam detection, providing a reliable solution that could significantly enhance the quality of user interactions and maintain the integrity of online platforms like YouTube.

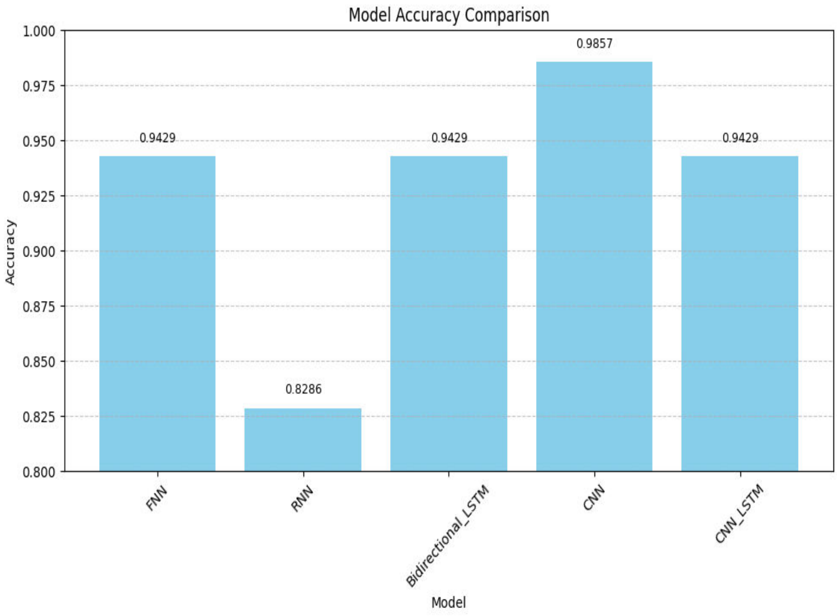


Fig. 3. Model Accuracy Comparison

- **Comparison with Previous Work:**

Table 2: Experimental results using the machine learning model for the YouTube02_KatyPerry dataset.

ML Approach	**Accuracy**	**Precision**	**F1**	**Recall**
Naïve Bayes	0.8571	0.8205	0.8648	0.9142
Logistic Regression	0.9142	0.9677	0.9090	0.8571
KNN	0.8540	0.8571	0.8421	0.8275
MLP	0.9297	0.9021	0.9273	0.9540
Random	**0.9459**	**0.9720**	**0.9404**	**0.9080**

Forest				
Decision Tree	0.9285	0.9285	0.9285	0.9285

In comparison to existing models in the literature, our CNN model's performance outshines several traditional classifiers from previous studies. For example, in the work [7], machine learning approaches such as Naïve Bayes achieved an accuracy of 85.71%, while Logistic Regression and K-Nearest Neighbors (KNN) reported accuracies of 91.42% and 85.40%, respectively. Additionally, more advanced models like Random Forest and MLP reached accuracies of 94.59% and 92.97%. In contrast, our CNN model demonstrated superior accuracy at 98.57%, with consistently high precision, recall, and F1-score values. This makes it a far more reliable solution for spam detection tasks.

Overall, the CNN model's outstanding performance highlights its potential as a valuable tool for spam detection on platforms like YouTube. Its ability to surpass traditional classifiers and leverage deep learning for pattern recognition demonstrates its strength in accurately identifying spam.

V. CONCLUSION

The deep learning-based spam detection model, utilizing a dataset of YouTube comments, demonstrated exceptional effectiveness, achieving an impressive accuracy of 98.57% with the Convolutional Neural Network (CNN) model outperforming other approaches. This high accuracy underscores the model's ability to accurately classify comments as spam or non-spam, which is vital for maintaining the integrity of online conversations and improving user experiences on social media platforms.

Among the individual models, CNN outperformed others, showing the highest accuracy and demonstrating its capability to capture complex patterns and contextual relationships within text data. The integration of CNN with other deep learning techniques, such as LSTM and Bidirectional LSTM, further highlighted the potential of combining multiple models to enhance spam detection performance, showing that leveraging the strengths of different algorithms yields better results than relying on a single approach.

This research emphasizes the significant role of advanced deep learning methods, particularly CNN-based models, in the detection of spam comments. It adds to the growing body of evidence supporting the use of machine learning in content moderation, paving the way for more accurate and efficient spam detection tools that can improve trust and safety in online communities. Future studies could explore additional datasets, including multilingual and domain-specific spam, implement attention mechanisms, and integrate ensemble learning techniques to further boost detection precision, adaptability, and reliability in real-world scenarios.

REFERENCES

[1] Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges. Security and Communication Networks, 2022, 1–19.

[2] Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., et al. (2022). Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. Mathematics, 10(22), 4173.

[3] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 1–24.

[4] Danilchenko, K., Segal, M., & Vilenchik, D. (2022). Opinion spam detection: A new approach using machine learning and network-based algorithms. In Proceedings of the International AAAI Conference on Web and social media, vol. 16 (pp. 125–134).

[5] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., et al. (2023). Scaling vision transformers to 22 billion parameters. In International Conference on Machine Learning (pp. 7480–7512). PMLR.

[6] Grewal, N., Nijhawan, R., & Mittal, A. (2022). Email spam detection using machine learning and feature optimization method. In Distributed Computing and Optimization Techniques: Select Proceedings of ICDCOT 2021 (pp. 435–447). Springer.
[7] Guo, Y., Mustafaoglu, Z., & Koundal, D. (2023). Spam detection using bidirectional transformers and machine learning classifier algorithms. Journal of Computational and Cognitive Engineering, 2(1), 5–9.
[8] James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). An Introduction to Statistical Learning (Vol. 112). Springer.
[9] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. ACM Computing Surveys (CSUR), 54(10s), 1–41.
[10] Mashaleh, A. S., Ibrahim, N. F. B., Al-Betar, M. A., Mustafa, H. M., & Yaseen, Q. M. (2022). Detecting spam email with machine learning optimized with Harris Hawks optimizer (HHO) algorithm. Procedia Computer Science, 201, 659–664.
[11] Yellapu, Jhansi, et al. "Spam Comment Detection Using the Ensemble Technique." 2023 4th International Conference on Intelligent Technologies (CONIT). IEEE, 2024.
[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.
[13] M.N Istiaq Ahsam, Tamzid Nahian,Abdullah All Kafi, Md. Ismail Hossain, Faisal Muhammad Shah, “Review Spam Detection using Active Learning” in Proc. of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)
[14] Goswami, A., Patel, R., Mavani, C., & Mistry, H. K. (2024). Identifying Online Spam Using Artificial Intelligence. International Journal on Recent and Innovation Trends in Computing and Communication, 12(2),
[15] Kumar, N., & Sonowal, S. (2020, July). Email spam detection using machine learning algorithms. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113).
[16] Sonare, B., Dharmale, G. J., Renapure, A., Khandelwal, H., & Narharshettiwar, S. (2023, May). E-mail Spam Detection Using Machine Learning. In 2023 4th International Conference for Emerging Technology.
[17] Mccord, M., & Chuah, M. (2011). Spam detection on twitter using traditional classifiers. In Autonomic and Trusted Computing: 8th International Conference, ATC 2011, Banff, Canada, September 2-4, 2011. Proceedings 8 (pp. 175-186).
[18] Kabakus, A. T., & Kara, R. (2017). A survey of spam detection methods on twitter. International Journal of Advanced Computer Science and Applications, 8(3).
[19] Shaik, C. M., Penumaka, N. M., Abbireddy, S. K., Kumar, V., & Aravinth, S. S. (2023, February). Bi-LSTM and conventional classifiers for email spam filtering. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 1350-1355).
[20] Airlangga, G. (2024). Spam Detection in YouTube Comments Using Deep Learning Models: A Comparative Study of MLP, CNN, LSTM, BiLSTM, GRU, and Attention Mechanisms. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4(4), 1533-1538.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details