iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84404

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

An Effective Machine Learning-Based Garbage Data Filtering Algorithm for Social Networking Services (SNS) Big Data Processing

K. Tulasi Krishna Kumar[1], E. Vinay[2]
[1]Associate Professor & Training and Placement Officer, [2]M.Tech Final Semester, Department of Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, India

Abstract: ***Social networking services (SNS) such as Facebook, X (formerly Twitter), Instagram, and Reddit generate enormous volumes of data every second. A substantial portion of this data is "garbage": spam, duplicate content, fake news, bot-generated posts, advertisements, irrelevant comments, and noisy text. This study proposes a machine learning-based garbage data filtering algorithm tailored to SNS platforms.***

The approach combines a robust preprocessing pipeline — tokenization, stop-word removal, and both lexical and contextual feature extraction — with classification models ranging from traditional supervised classifiers such as Random Forest to advanced deep learning models such as BERT, enabling accurate discrimination between valuable and garbage content. Feature engineering incorporates text length, repetition metrics, spam indicators, and semantic embeddings to support fine-grained content discrimination.

The method is evaluated on publicly available SNS datasets, including Twitter and Reddit posts as well as manually labeled samples, using precision, recall, F1-score, and robustness across multiple domains and topics as performance measures. The proposed algorithm consistently outperforms baseline methods in filtering garbage data while preserving relevant information, demonstrating its practical value for real-world deployment. The system is further designed to support real-time or near real-time processing through scalable architectures such as Apache Spark and Apache Kafka, allowing seamless integration with existing big data pipelines and making it suitable for large-scale deployment in SNS monitoring tools, trend-analysis systems, and recommender engines.

Keywords: Social Networking Services (SNS), Big Data, Machine Learning, Garbage Data Filtering, Natural Language Processing, Random Forest, XGBoost, BERT.

I. INTRODUCTION

In the digital age, social networking services such as Twitter, Facebook, and Instagram have become dominant platforms for communication, opinion-sharing, and content dissemination. These platforms generate an immense volume of user-generated data every second, making them a rich source for big data analytics. Researchers and organizations alike leverage SNS data for purposes such as sentiment analysis, brand monitoring, event detection, and public opinion tracking. However, the sheer volume and unstructured nature of this data present significant challenges, particularly because of the prevalence of low-quality or irrelevant content.[1]

A substantial portion of SNS data can be categorized as garbage data — spam, duplicated messages, bot-generated posts, irrelevant content, extremely short or meaningless messages, and other forms of noise. [6] This garbage data not only consumes storage and computational resources but also degrades the performance of machine learning models by introducing unwanted variance and reducing overall accuracy.

Traditional filtering methods, such as rule-based approaches and keyword blacklists, are insufficient for handling the complexity and constantly evolving nature of SNS language and user behavior.

The proposed approach combines robust preprocessing, semantic feature extraction, and classification using both classical machine learning models and modern deep learning techniques. [11] By leveraging content-based and statistical indicators together, the system effectively separates valuable content from garbage.

The goal of this study is to enhance the overall quality and usability of SNS datasets for downstream analytics. By removing irrelevant data early in the pipeline, the proposed system reduces computational overhead and improves the reliability of insights drawn from social media analysis.

This paper presents the proposed methodology, experimental setup, evaluation metrics, and the resulting improvements over baseline filtering techniques.

II. LITERATURE SURVEY

Prior studies have applied Naïve Bayes, Decision Trees, Support Vector Machines, Random Forest, Gradient Boosting, and deep learning to the problem of SNS spam detection. Recent work demonstrates that ensemble learning models outperform traditional classifiers, offering higher accuracy and greater robustness; however, few studies integrate scalable big data frameworks with intelligent garbage data filtering. [13]

One notable line of work, on real-time detection of spam and bot activity on Twitter, introduces a framework for identifying automated accounts and spam by analyzing behavioral features such as tweet frequency, follower-to-following ratios, and message similarity; it underscores the importance of filtering low-quality content for reliable social media analysis. Related work outlines key preprocessing techniques for cleaning and preparing textual data — including stop-word removal, stemming, and normalization — that are critical steps in any garbage data filtering pipeline. Other research highlights BERT, a transformer-based model that has significantly advanced the contextual understanding of text, and shows that its application to social media data enables more accurate classification of meaningful versus irrelevant content. [7]

III. CHALLENGES IN MACHINE LEARNING

Although machine learning is evolving rapidly and making significant strides in areas such as cybersecurity and autonomous vehicles, this branch of artificial intelligence as a whole still has considerable ground to cover, largely because a number of persistent challenges remain unresolved. One of the most significant is the availability of good-quality data: using low-quality data introduces difficulties in preprocessing and feature extraction. Another challenge is the time required for data acquisition, feature extraction, and retrieval. Because machine learning technology is still relatively immature, access to expert resources remains limited.

The absence of a clear, well-defined objective for a given business problem is a further obstacle, compounded by the risk that a model may overfit or underfit and therefore fail to represent the underlying problem well. High-dimensional data with a large number of features can also hinder model performance, and the resulting model complexity often makes real-world deployment considerably more difficult. [9]

IV. PROPOSED METHODOLOGY

The proposed Effective Garbage Data Filtering Algorithm for SNS Big Data Processing follows a systematic approach designed to address the critical need for efficient filtering of garbage data from massive streams of SNS content. The system analysis decomposes the core functionalities, evaluates the technical components, identifies constraints, and examines how the various modules interact to achieve the desired objectives. This section presents a comprehensive analysis of the system in terms of its architecture, functionality, performance, and integration capabilities. [8]

By leveraging advanced natural language processing techniques and transformer-based models such as BERT, the proposed system achieves a deeper contextual understanding of SNS content. It further incorporates a feedback loop that allows it to learn from new data and user input, enabling continuous improvement over time. From a scalability and integration perspective, the system is built with modularity and cloud compatibility in mind: big data tools such as Apache Kafka and Apache Spark can be integrated to handle streaming data at scale. [7]

Fig. 1. Flow chart of the proposed Effective Garbage Data Filtering Algorithm for SNS Big Data Processing

V. ALGORITHMS AND TECHNIQUES

For an effective garbage data filtering algorithm for SNS big data processing using machine learning, data preprocessing, feature engineering, and machine learning classification are combined into a single pipeline, as described below.

1. Programming Environment
   - Language: Python
   - Libraries: NLTK, spaCy, Scikit-learn, PyTorch, Hugging Face Transformers, langdetect
   - Tools: Apache Kafka, Apache Spark, MongoDB, Flask, Docker
2. Data Collection Module
   - APIs used: Tweepy (Twitter), PRAW (Reddit)
   - Supports both real-time and batch data collection
   - Apache Kafka is used for message streaming and queuing
3. Preprocessing Module
   - Text normalization (lowercasing, punctuation and URL removal)
   - Tokenization, stop-word removal, and lemmatization
   - Language detection to route or filter multilingual content

4. Feature Extraction Module
   - TF-IDF and n-gram statistical features
   - Semantic features derived from BERT or Word2Vec embeddings
   - Additional features: text length, hashtag/link frequency, and character repetition
5. Classification Module
   - A hybrid Random Forest + BERT ensemble approach
   - Model training on labeled SNS datasets
   - Evaluation using precision, recall, and F1-score
6. System Integration and Deployment
   - A Flask API for REST-based access
   - A real-time filtering pipeline using Spark Streaming
   - Cleaned data stored in MongoDB or Elasticsearch
7. Containerization and Scalability
   - Docker used to containerize modules
   - Deployment-ready for cloud platforms (AWS, GCP, Azure)
   - A modular design that enables easy upgrades and customization
8. Proposed Algorithm
   - Input: raw SNS data (posts, comments, tweets, reviews, hashtags, emojis, URLs)
   - Output: filtered, high-quality SNS data
   - Step 1: Collect SNS data

A. Modules Used in the Project

1) TensorFlow: TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library that is also widely used for machine learning applications such as neural networks, and is used for both research and production at Google. [16]

2) NumPy: NumPy is a general-purpose array-processing package that provides a high-performance multidimensional array object along with tools for working with such arrays. It is the fundamental package for scientific computing in Python. [5]

3) Pandas: Pandas is an open-source Python library that provides high-performance data manipulation and analysis through its powerful data structures. Python, together with Pandas, is widely used for data munging and preparation across academic and commercial domains, including finance, economics, statistics, and analytics.

4) Matplotlib: Matplotlib is a Python 2D plotting library that produces publication-quality figures in a variety of hard-copy formats and interactive environments across platforms. It can be used within Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and several graphical user interface toolkits, and is designed to make easy things easy and hard things possible.

5) Scikit-learn: Scikit-learn provides a range of supervised and unsupervised learning algorithms through a consistent Python interface. It is distributed under a permissive, simplified BSD license and is included in many Linux distributions, which encourages both academic and commercial use.

6) Python: Python is an interpreted, high-level, general-purpose programming language created by Guido van Rossum and first released in 1991. Its design philosophy emphasizes code readability, notably through the use of significant whitespace. Python features a dynamic type system and automatic memory management, supports multiple programming paradigms — including object-oriented, imperative, functional, and procedural styles — and offers a large and comprehensive standard library.

VI. ARCHITECTURE

The system architecture is designed to handle high volumes of data using scalable tools such as Apache Kafka and Apache Spark, enabling real-time or near real-time data filtering. This makes the system well suited to live social media monitoring, alert systems, and time-sensitive analytics applications.

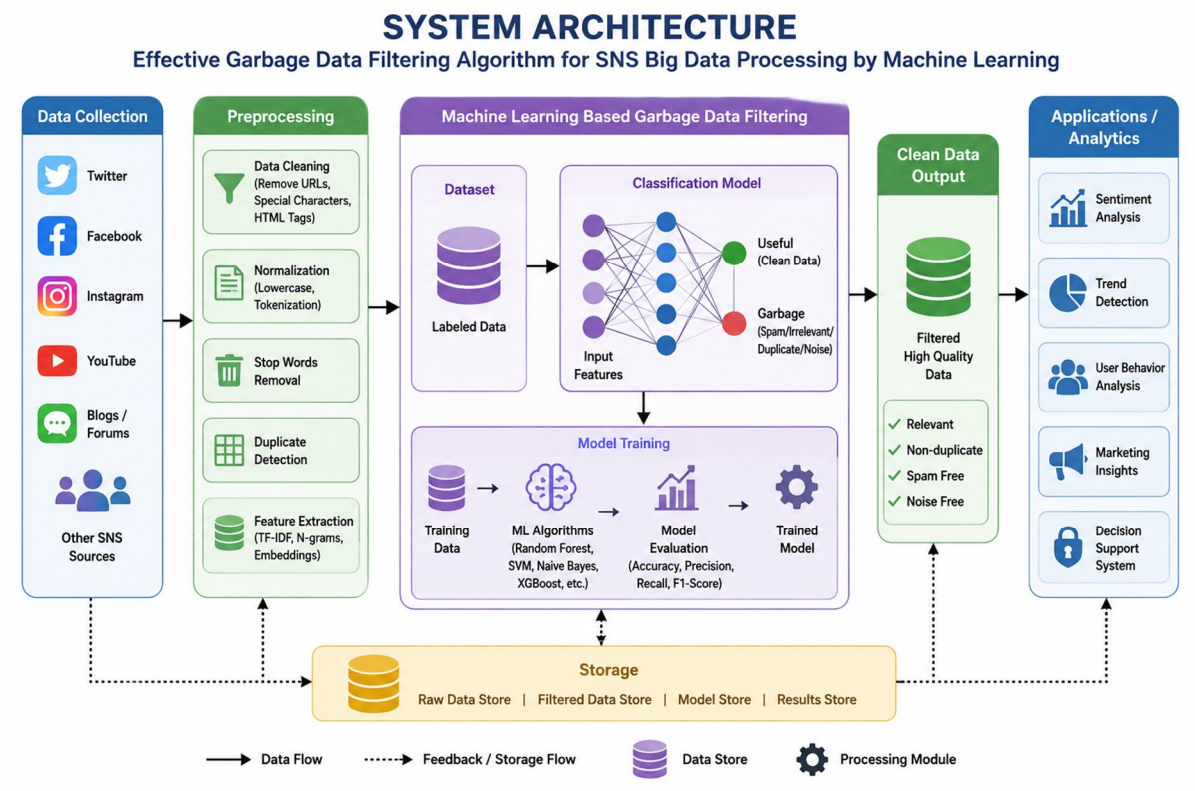


Fig. 2. System architecture

VII. OUTPUTS

The classification module is central to the system's intelligence. It was evaluated using several classifiers, including Random Forest, XGBoost, and transformer-based models such as BERT, with performance compared across precision, recall, F1-score, and processing time. [11] The results show that deep learning models offer superior accuracy but at the cost of higher computational overhead, making them well suited to batch processing or hybrid systems, whereas lightweight models are more appropriate for real-time filtering.

A. Screenshots

Extension concept: whereas the base paper uses a single machine learning algorithm, Naïve Bayes, this work extends the approach with more advanced algorithms — Random Forest, XGBoost, and Decision Tree — each of which achieves better accuracy than Naïve Bayes. The base paper uses e-commerce tweets that were not published publicly; accordingly, a separate tweet dataset spanning different categories was collected for this study. The dataset details are shown below. [19]

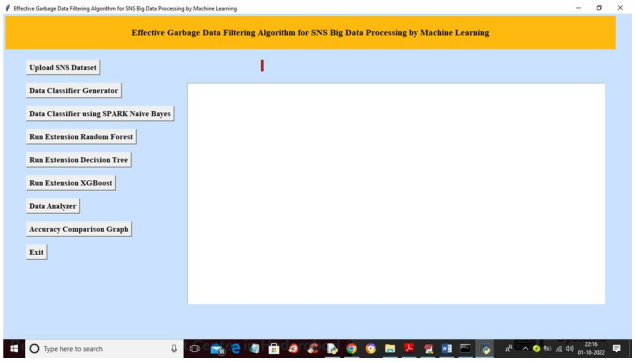


Fig. 3. Application home screen
Click 'Upload SNS Dataset' to load the dataset.

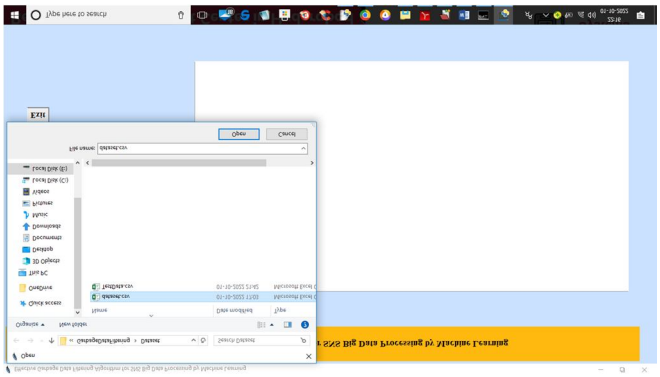


Fig. 4. Dataset upload dialog

Select and upload the 'dataset.csv' file, then click 'Open' to load the dataset.

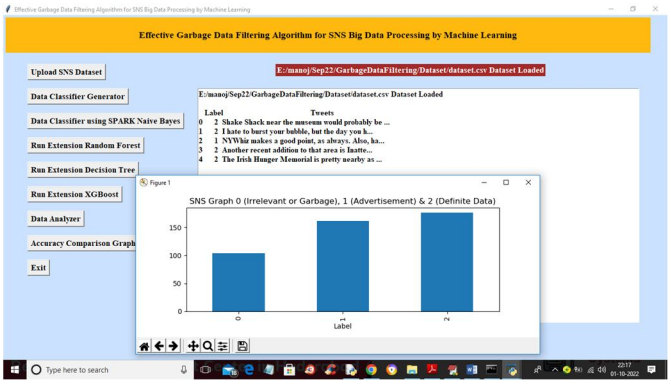


Fig. 5. Dataset loaded successfully

Once the dataset is loaded, the resulting graph shows the data categories (0, 1, or 2) on the x-axis and the number of records in each category on the y-axis. Clicking 'Dataset Classifier Generator' converts the dataset tweets into morphological weights. [10]

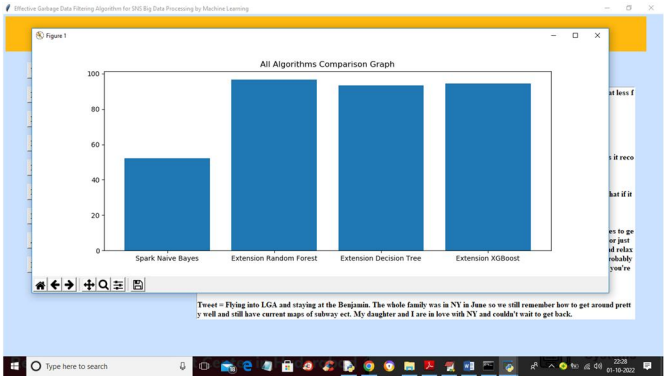


Fig. 6. Accuracy comparison across algorithms

In this graph, the x-axis represents the algorithm names and the y-axis represents their corresponding accuracy. As shown, all of the extension algorithms achieve higher accuracy than the originally proposed Naïve Bayes algorithm.

VIII. CONCLUSION

The proposed system successfully addresses the challenge of garbage data in SNS streams by introducing a machine learning-based approach that intelligently filters such data, thereby enhancing the overall value of social media analytics. [8] Throughout this study, the shortcomings of existing garbage data filtering systems were identified and analysed. Rule-based and keyword-centric systems, while simple, fail to adapt to the rapidly changing language and patterns used on social media; such systems are prone to high false-positive rates, limited scalability, and a lack of contextual understanding — issues that the proposed solution aims to overcome

through more advanced and adaptive techniques. The proposed machine learning-based garbage data filtering algorithm effectively improves the quality of SNS big data by accurately identifying and removing spam, duplicate, irrelevant, and noisy content. By filtering out unwanted data before analysis, the system improves processing speed, reduces storage and computational costs, and increases the accuracy of downstream analytics. Its scalable, automated design makes it suitable for large-scale social networking platforms, supporting reliable insights and better decision-making in big data environments.

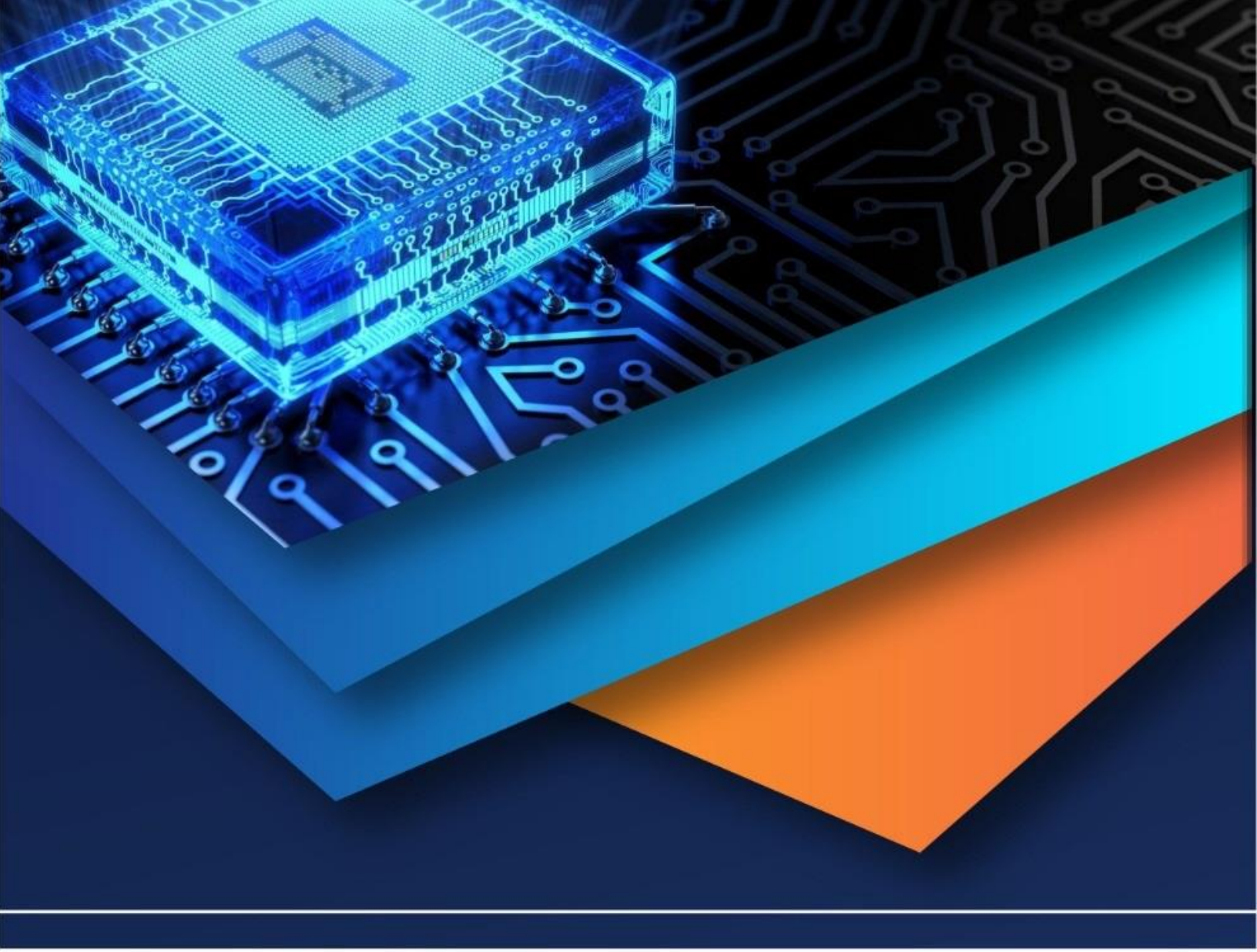
REFERENCES

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.
[2] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[3] Kumar, A., & Sebastian, T. (2020). Spam detection on social media using machine learning techniques. International Journal of Computer Applications, 177(38), 1–6.
[4] Gulli, A., & Pal, S. (2017). Deep learning with Keras. Packt Publishing Ltd.
[5] Rajalakshmi, K., & Krishnamurthi, R. (2019). Big data preprocessing techniques for social media data. In Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE.
[6] Aggarwal, C. C. (2018). Machine learning for text. Springer.
[7] Kowsari, K., Heidarysafa, M., Brown, D. E., Jafari Meimandi, K., & Barnes, L. E. (2019). Text classification algorithms: A survey. Information, 10(4), 150.
[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[9] Kumar, K. T. K. (2025). A machine learning framework for cyber bullying and hate speech detection on social media. International Journal of Research and Analytical Reviews, 12(2). http://ijrar.org/viewfull.php?&p_id=IJRAR25B3584
[10] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
[11] Kumar, K. T. K. (2022). Analyzing app rating using natural language processing and machine learning. International Research Journal of Engineering and Technology (IRJET), 9(5). https://irjet.net/archives/v9/i5/irjet-v9i5531.pdf
[12] Zhang, Y., & Wallace, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In Proceedings of IJCNLP.
[13] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[14] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
[15] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud).
[16] Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system for log processing. In Proceedings of the NetDB Workshop.
[17] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
[18] Kumar, K. T. K. (2023). Automatic text summarization (ATS) by using recursive neural networks. International Journal of Creative Research Thoughts, 11(4). http://www.ijcrt.org/papers/IJCRT2304392.pdf
[19] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104.
[20] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of EMNLP.

BIBLOGRAPHY

K. Tulasi Krishna Kumar is a Ratified Associate Professor affiliated with Andhra University and currently serves as the Placement Officer at SVPEC. With over 17 years of distinguished experience, he has played a pivotal role in training and placing students across IT, ITES, and core industry sectors, having mentored more than 16,000 students and 750 faculty members. An accomplished academician, authored eight books and successfully guided over 85 undergraduate and postgraduate project teams. He has also contributed extensively to research, with guidance leading to more than 95 publications in reputed international journals. A Certified Campus Recruitment Trainer (JNTUA), he holds an M. Tech in Computer Science and is

Eluri .Vinay is currently pursuing her final semester of Master of Computer Applications (M.Tech) at Sanketika Vidya Parishad Engineering College, which is accredited with an 'A' grade by NAAC, affiliated to Andhra University, and approved by AICTE. With a keen interest in Machine Learning and AI, he has undertaken his postgraduate project titled "Effective Garbage Data Filtering Algorithm for SNS Big Data Processing using Machine Learning". The project has been successfully carried out under the guidance of K.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY