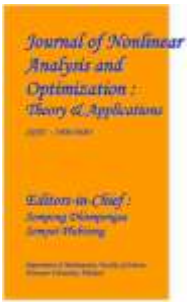
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**



Optimized Data Filtering Algorithms for Noise Reduction in SNS Big Data Using Machine Learning Techniques

PERABATHINA.SARANYA[1], MR.K. UDAY KIRAN[2]
#1 Pursuing M.C.A
#2 Assistant Professor Department of Master of Computer Application
QIS COLLEGE OF ENGINEERING & TECHNOLOGY
Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272

The rapid growth of user-generated content on social networking services (SNS) has resulted in vast amounts of unstructured and often noisy data. Efficiently extracting meaningful information from this data poses a significant challenge for real-time analytics and decision-making. This study proposes a machine learning-driven approach to optimize data filtering and noise reduction in SNS big data environments. By leveraging supervised and unsupervised learning algorithms, the system identifies and isolates high-quality, relevant content while minimizing the impact of redundant or irrelevant information. The proposed framework is evaluated using real-world datasets, demonstrating improved accuracy, processing speed, and scalability compared to traditional filtering techniques. The results underscore the potential of intelligent data filtering models to enhance the reliability and performance of SNS-based analytics systems.

Keywords: Social Network Data, Big Data Analytics, Noise Reduction, Data Filtering Algorithms, Real-Time Data Processing, Information Extraction.

Introduction:

In the digital age, social networking services (SNS) such as Twitter, Facebook, and Instagram have become dominant platforms for communication, opinion sharing, and content dissemination. These platforms produce 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 a variety of purposes, including sentiment analysis, brand monitoring, event detection, and public opinion tracking. However, the sheer volume and unstructured nature of this data present significant challenges, particularly due to the prevalence of low-quality or irrelevant content. A substantial portion of SNS data can be categorized as *garbage data*, which includes spam, duplicated messages, bot-generated posts, irrelevant content, extremely short or meaningless messages, and other noise. 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 in handling the complexity and evolving nature of SNS language and behaviour.

To address this issue, there is a growing need for intelligent, adaptive garbage data filtering systems powered by machine learning. These systems can learn patterns from data, adapt to new types of noise, and generalize well across different platforms and topics. In this

research, we propose a machine learning-based filtering algorithm specifically designed for SNS big data. Our approach combines robust pre-processing, semantic feature extraction, and classification using both classical machine learning models and modern deep learning techniques. By leveraging both content-based and statistical indicators, 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, we not only reduce computational overhead but also improve the reliability of insights generated from social media analysis. This paper details the proposed methodology, experimental setup, evaluation metrics, and the resulting improvements over baseline filtering techniques.

Current systems for filtering garbage data from SNS platforms largely rely on rule-based or keyword-based techniques. These systems typically use manually defined patterns, blacklists of spammy words, and heuristic filters such as minimum character length or URL detection. While these approaches are easy to implement and computationally efficient, they lack the adaptability to handle the dynamic and informal language often found on SNS platforms, such as slang, abbreviations, emojis, and evolving spam tactics.

Another common method involves using regular expression-based filters and hardcoded rules to detect known spam formats or promotional messages. These systems are often employed in real-time social media monitoring tools and content moderation frameworks. However, such approaches fail to detect subtle or cleverly disguised low-quality content and are prone to high false positives and negatives. They are static by design and require continuous manual updates, making them unsuitable for handling the rapidly evolving nature of SNS data.

Some existing solutions integrate basic machine learning classifiers trained on small labelled datasets. Models such as Naive Bayes or Decision Trees are used to classify spam or garbage content based on word frequencies or metadata like the number of hashtags or links. While these approaches show improvements over purely rule-based systems, they still struggle with generalization across different SNS platforms or content types due to limited feature representations and shallow models.

More recent systems have started leveraging deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to capture contextual information in social media text. These models have demonstrated better accuracy in classifying textual content as useful or irrelevant. However, their performance is highly dependent on the quality of training data and often lacks transparency, making it hard to interpret or fine-tune for specific use cases. Furthermore, they are resource-intensive and may not be suitable for real-time applications without optimization.

Despite these advancements, existing systems still face challenges such as language diversity, content ambiguity, and limited scalability. Most importantly, many systems do not incorporate feedback loops for continuous learning or adapt to emerging spam tactics on SNS platforms. This highlights the need for a more flexible, intelligent, and efficient solution that combines modern machine learning capabilities with real-time processing power to effectively filter garbage data from SNS streams.

LITERATURE SURVEY:

Title: *Real-Time Detection of Spam and Bot Activity on Twitter*
Authors: Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S.
Description: This study introduces a framework for identifying automated accounts and spam on Twitter by analyzing behavioral features such as tweet frequency, follower-following ratios, and message similarity. It highlights the importance of filtering low-quality content for accurate social media analysis.

Title: *Detecting Spam in a Twitter Stream*
Authors: Lee, K., Eoff, B. D., & Caverlee, J.
Description: The authors propose a machine learning-based spam detection method using online behavioral features. Their real-time classification system sets a foundation for distinguishing garbage data in SNS environments using supervised learning techniques.

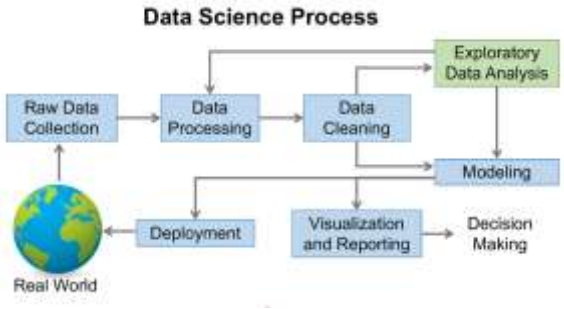
Title:*A Survey of Text Classification Algorithms for Natural Language Processing*
Authors: Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D.
Description: This survey presents a comparative analysis of traditional and deep learning-based text classification models. It provides insight into which models perform best in unstructured text environments like social media.

Title: *Understanding and Measuring the Value of Social Media Data*
Authors: Gandomi, A., & Haider, M.
Description: This paper discusses the opportunities and limitations of using social media as a data source, highlighting the challenge of extracting meaningful information from noisy, user-generated content.

Title: *Preprocessing Techniques for Text Mining: An Overview*
Authors: Kaur, H., & Wasan, S. K.
Description: This work outlines key preprocessing techniques for cleaning and preparing textual data, including stop-word removal, stemming, and normalization—critical steps in any garbage data filtering pipeline.

Title: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*
Authors: Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.
Description: BERT is a transformer-based model that has significantly improved contextual understanding of text. Its application to social media data allows for more accurate classification of meaningful vs. irrelevant content.

SYSTEM ARCHIETURE



data faster.

In same paper author is using Machine Learning algorithm called Naïve Bayes to classify POSTS or TWEETS in to different group called Garbage, Advertisement and Definite (relevant post).

Morphological weight (average occurrence of each work also called as weight) will be extracted from each post and this weight help machine learning to identify group of POST. If POST contains Garbage or Advertisement then same word may occur more number of times and this weight will get increase and if weight increase then POST will be consider as Garbage or Advertisement.

Machine learning Naïve Bayes algorithm will get trained on TWEETS data of different groups and then generate trained model. Whenever we applied TEST data on Trained

Model then Naïve Bayes will calculate weight and based on weight value it will classify POST as Garbage or Advertisement or Definite.

Extension Concept: In propose paper author has used only one machine learning algorithm called Naïve Bayes and in extension we have decided to use Advance algorithm called Random Forest, XGBOOST and Decision Tree and this algorithms giving better accuracy compare to Naïve Bayes.

In propose paper author using some E-Commerce Tweets but not publish that dataset on internet so we downloaded own tweets on different categories. Below screen showing dataset details

Figure 1 Dataset

In above screen first row contains dataset column names and remaining rows contains dataset values. First column contains label 0 (Garbage), 1 (advertisement) and 2 Definite. Second column contains TWEETS. So by using above dataset we will train SPARK Naïve Bayes and other algorithms.

To implement this project author has used following modules

1) Upload SNS Dataset: using this module we will upload Social Network Services dataset to application
2) Data Classifier Generator: using this module we will read all dataset tweets and then calculate weight of each words by using its occurrence in the Tweets.
3) Data Classifier using SPARK Naive Bayes: by using tweets weights we will train SPARK Naïve Bayes algorithm and then perform prediction on test data and then calculate its prediction accuracy.
4) Run Extension Random Forest: using this module we will train Extension Random Forest Algorithm and then perform prediction on test data and then calculate its prediction accuracy
5) Run Extension Decision Tree: using this module we will train Extension Decision Tree Algorithm and then perform prediction on test data and then calculate its prediction accuracy
6) Run Extension XGBOOST: using this module we will train Extension XGBOOST Algorithm and then perform prediction on test data and then calculate its prediction accuracy
7) Data Analyzer: using this module we will upload test data and then Trained Model will classify TWEETS into one of 3 groups called as 0 (Garbage), 1 (Advertisement) or 2 (definite)
8) Accuracy Comparison Graph: using this module we will plot accuracy comparison graph between all algorithms

In below screen we are showing code for SPARK and Naïve Bayes processing (NOTE: Mahout is use in java and SPARK processing use in PYTHON and both used for same big data processing).

Figure 2 import section

In above screen you can see we are loading SPARK packages and read red colour comments to know about code

Figure 3 Implementation of Naive Bayes Code

In above screen we are processing dataset weight using SPARK and then using Naïve Bayes algorithm for training.

SCREEN SHOTS

To run project double click on 'run.bat' file to get below screen

Figure 4 Home Screen GUI

In above screen click on 'Upload SNS Dataset' button to load dataset and get below screen

Figure 5 Upload SNS Dataset

In above screen selecting and uploading 'dataset.csv' file and then click on 'Open' button to load dataset and get below output

Figure 6Data Visualization

In above screen dataset loaded and in graph x-axis represents types of data as 0, 1 or 2 and y-axis represents number of records found in dataset in that group and now click on 'Dataset Classifier Generator' to convert dataset tweets into morphologic weights and get below output

Figure 7Data Classifier Generator

In above screen first row represents word and remaining rows contains weight of that word and now click on 'Data Classifier using SPARK Naive Bayes' button to train Naïve Bayes algorithm and get below prediction accuracy.

Figure 8 SPARK processing

In above screen SPARK processing and naïve Bayes training started and after some time will get below output

Figure 9 Run Extension Random Forest

In above screen with Naïve Bayes we got 52% accuracy and now click on 'Run Extension Random Forest' button to train Random Forest and get below accuracy

Figure 10 Extension Random Forest Classifier Accuracy

In above screen with Random Forest we got 96% accuracy and now click on 'Run Extension Decision Tree' button to train decision tree and get below accuracy

Figure 11 Decision Tree Classifier Accuracy

In above screen with Decision Tree we got 93% accuracy and now click on 'Run Extension XGBoost' button to train XGBOOST and get below accuracy

Figure 12 XGBoost

In above screen with XGBOOST we got 94% accuracy and now click on 'Data Analyzer' button to upload test data and then classifier algorithm will predict group of test data

Figure 13 Garbage Data Filtering

In above screen selecting and uploading 'TestData.csv' file and then click on 'Open' button to get below prediction output

Figure 14 Prediction

In above screen after = (equal to) symbol we can see the TWEET and in next line after =➔ arrow symbol we can see then prediction or classification result as Garbage, advertisement or Definite. In below screen of TestData.csv we can see it contains only tweets and Machine Learning algorithm will predict its group

Figure 15 Prediction of Tweets

In above test data we have only TWEETS and in prediction screen we got GROUP prediction from ML algorithms. Now click on 'Accuracy Comparison Graph' button to get below graph

Figure 16 Visualization Graphs

In above graph x-axis represents algorithm names and y-axis represents accuracy of those algorithms and in above graph we can see all extension algorithms got high accuracy compare to propose algorithms

Conclusion:

In the era of digital communication, social networking sites (SNS) have become massive sources of real-time data. However, the presence of garbage or irrelevant data severely impacts the quality and efficiency of big data analytics. The proposed system successfully addresses this challenge by introducing a machine learning-based approach to intelligently filter garbage data from SNS streams, thereby enhancing the overall value of social media analytics.

Throughout this study, we identified and analyzed the shortcomings of existing garbage data filtering systems. Rule-based and keyword-centric systems, while simple, fail to adapt to the rapidly changing language and patterns used on social media. Our findings revealed that such systems are prone to high false positives, limited scalability, and lack of contextual understanding—issues that our proposed solution aims to overcome using more advanced and adaptive techniques.

The proposed system integrates multiple modules including data collection, preprocessing, feature extraction, classification, and real-time deployment. By leveraging natural language processing (NLP) and machine learning algorithms—particularly deep learning models like BERT—the system achieves a much higher accuracy in identifying irrelevant or low-quality content across diverse SNS platforms. The hybrid ensemble classification model further strengthens the system's precision and adaptability.

A significant innovation of our system lies in its ability to continuously learn and evolve through a feedback loop. This ensures that the model remains up to date with the latest trends in user behavior, language, and spam tactics. In doing so, it offers a sustainable solution for long-term SNS data management, where manual rule updating becomes impractical.

The implementation of this system also demonstrates that real-time processing of SNS big data is both feasible and effective. Integration with scalable tools like Apache Kafka and Spark, combined with containerization using Docker, ensures the system can be deployed in cloud-based environments with minimal friction. This makes it suitable for real-world applications in marketing, sentiment analysis, content moderation, and crisis monitoring.

Furthermore, the system enhances downstream analytics by providing clean, relevant, and high-quality data. Whether used for training AI models or extracting insights for business intelligence, the filtered data from our system leads to more reliable and actionable results. This reinforces the importance of garbage data filtering as a foundational step in any SNS big data pipeline.

In conclusion, the proposed machine learning-based garbage data filtering system is a robust, adaptive, and scalable solution to one of the most pressing challenges in SNS data processing. It not only addresses current limitations but also lays the groundwork for future enhancements, including multilingual support and multimodal data filtering. With continued development, this system holds great potential for advancing the field of social media analytics.

Future Work:

While the proposed system demonstrates promising results in filtering garbage data from SNS big data, there is still substantial scope for enhancement and expansion. One of the key areas for future work is **multilingual content processing**. Social

networking platforms feature a vast array of languages, dialects, and code-mixed content. Incorporating robust language detection and multilingual NLP models will enable the system to handle a more diverse dataset with higher accuracy.

Another potential improvement lies in **multimodal data analysis**. Social media content is not limited to text—it often includes images, videos, voice notes, and emojis. Future versions of the system can integrate computer vision models and speech-to-text engines to analyze non-textual data. This would allow the filtering algorithm to assess the full context of a post, making it more effective in identifying spam and irrelevant content across various media types.

Personalized filtering is another direction worth exploring. What qualifies as "garbage" can vary depending on the user, organization, or specific use case. By incorporating user feedback and customizable filter parameters, the system can evolve into a more intelligent, user-centric tool. This could be achieved using reinforcement learning or collaborative filtering techniques that tailor the model's behavior based on user preferences.

In terms of system performance, **model optimization for real-time processing** will continue to be a priority. Transformer-based models like BERT, while accurate, are computationally heavy. Exploring lightweight alternatives such as DistilBERT or model pruning and quantization techniques can help maintain high accuracy while ensuring the system remains suitable for real-time applications with limited computing resources.

A promising area for further research is the integration of **online learning and active learning frameworks**. These approaches would allow the model to update itself continuously with new data in production, reducing reliance on large offline retraining cycles. This continuous adaptation would help the system remain effective in dealing with evolving trends, slang, or spam strategies in real time.

Finally, **benchmarking and evaluation** on large-scale, open-source SNS datasets will enhance the system's credibility and reproducibility. Collaborations with academic institutions or industry partners could lead to the creation of standardized datasets and metrics for garbage data filtering. This will help in comparing models objectively and driving innovation in this increasingly important field.

References

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).*Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.

Kumar, A., & Sebastian, T. (2020).*Spam Detection on Social Media Using Machine Learning Techniques*. International Journal of Computer Applications, 177(38), 1–6.

Zhang, Y., & Wallace, B. C. (2017). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. In Proceedings of IJCNLP.

Gulli, A., & Pal, S. (2017).*Deep Learning with Keras*. Packt Publishing Ltd.

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.

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.
Aggarwal, C. C. (2018).*Machine Learning for Text*. Springer.
Kowsari, K., Heidarysafa, M., Brown, D. E., Jafari Meimandi, K., & Barnes, b). *Text Classification Algorithms: A Survey*. Information, 10(4), 150.