Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 29 (2026), No. 7 & 8, pp. 743–785
DOI : 10.47974/JSMS-1585

Noise elimination techniques in social media big data : A survey

Amitava Sarder *
School of Computer Science
Swami Vivekananda University
Barrackpore 700121
West Bengal
India

Ranjan Kumar Mondal †
Department of Computer Science and Engineering
Swami Vivekananda University
Barrackpore 700121
West Bengal
India

Abstract

In the era of big data, social media platforms have become invaluable sources of information, offering vast amounts of data that can be analyzed to extract valuable insights. With the explosive growth of social media and the large volume of data it generates, noise has become a significant challenge in deriving meaningful insights. This article provides a comprehensive survey of noise elimination techniques in social media big data.

In this paper, we review the literature on various noise elimination methods in social media big data and introduce a new classification of these strategies, such as preprocessing techniques, filtering methods, topic modelling, machine learning approaches, community detection, user-based filtering, outlier detection, anomaly detection, spam detection and related approaches. The survey examines the latest developments and recent advancements in noise reduction techniques, highlighting subcategories within each approach and their contributions to noise removal. Additionally, it discusses the challenges and opportunities associated with noise elimination in social media big data. This survey serves as a valuable resource for researchers and practitioners seeking an overview of noise reduction methods to improve the quality and reliability of social media big data analysis.

* *ORCID-ID:* **0000-0002-6210-5583**
† *ORCID-ID:* **0000-0002-4866-4004**

* *E-mail:* **sarder.amitava@gmail.com (Corresponding Author)**
† *E-mail:* **ranjankm@svu.ac.in**

Overall, this survey offers valuable insights into the landscape of noise elimination techniques in social media big data analysis, providing researchers and practitioners with a comprehensive understanding of existing methods and guiding future research in this important field.

Subject Classification: *97P80 Artificial intelligence (educational aspects).*

Keywords: *Social media, Big data, Noise elimination, Machine learning, Filtering methods.*

Introduction

Social media platforms have become vast collections of user-generated content, providing valuable insights into individuals' thoughts, choices and actions. However, the sheer volume of social media data presents a significant obstacle known as "noise." The rapid creation and collection of large amounts of social data, often called Big Social Data (BSD), pose major challenges because of the presence of noise [1]. Noise includes irrelevant, repetitive or misleading information that hinders the extraction of meaningful insights from the large pool of social media data. This noise can stem from various sources, such as spam, ads, fake accounts, troublemakers and trivial or duplicated posts. Recent work shows that online platforms boost both genuine updates and misleading content. They also allow certain users to push false claims. Methods inspired by natural systems can identify who shares real content and who spreads false claims. This makes later filtering and tracking tools more accurate [16]. Research shows that influencers have a big impact on buying decisions. Their posts often follow marketing patterns. Sometimes these posts look like genuine interest and other times like spam. That's why influencer content needs extra care when being filtered or labelled [8]. Noise makes it harder for machine learning algorithms to accurately detect and learn genuine patterns in the data, which can result in less accurate models and predictions. Social media data is often used for advertising and checking brand reputation. User actions, post style, and influencer posts all affect how people see a brand. These factors increase attention. They also create platform-specific noise that makes analysis harder [10]. Therefore, implementing effective noise removal techniques is crucial to maintain the integrity and reliability of social media big data analysis [2].

We mention a few recent papers that have reviewed the noise elimination strategies and mechanisms in social media big data:

- It has been demonstrated that incorporating efficient methods for negation handling, word n-grams and mutual information-based Feature Selection (F.S) can effectively remove noise and boost the performance of the Naïve Bayes classifier [3]. By combining these techniques, the study achieved a significant improvement in accuracy. Naïve Bayes classifiers were selected because of their speed, scalability, and robustness to noise and resistance to overfitting. The proposed supervised sentiment classification model can be applied to various text categorization tasks to improve both speed and accuracy. However, the research does not address unresolved issues or explore future areas of interest for researchers [3].
- It has been found that some researchers in their study focus on noise in the class label of the target attribute and suggest straightforward strategies for detecting and eliminating class noise in classification problems with improved performance. In cases of high noise levels, classifiers in ensembles tend to learn incorrect patterns from noisy data, leading to reduced precision and recall. Application of noise pre-processing techniques results in enhanced accuracy in identifying noisy data, thereby improving classification performance. The scope for future work in the article under consideration will explore determining optimal ensemble sizes, conducting in-depth analyses of noise handling methods and devising strategies specifically tailored for multiclass datasets [38].
- Others address the issue of noise in classification tasks and propose a new iterative noise filtering method that combines three different paradigms: ensembles, iterative filtering and noise measures to improve the accuracy of classification algorithms after filtering. The method uses the fusion of predictions from multiple classifiers and introduces a noisy score to control filtering sensitivity, allowing customization of the number of noisy examples removed in each iteration. It is compared against existing filters on real-world datasets with varying levels of class noise. To regulate the filter's sensitivity to noise, a noisy score is introduced. The findings show that the proposed method effectively balances removing noisy examples and preserving clean ones, outperforming other noise filters. Compared to alternative noise filters, their approach removes fewer clean examples. The method is extensively evaluated and compared with state-of-the-art techniques using datasets with class noise, highlighting its effectiveness across classifiers with different

noise sensitivities. For proper functioning, a good balance between eliminating noisy examples and accurately detecting noise-free ones must be maintained [4].

- Some authors in their paper empirically evaluate three filtering strategies—keyword filtering, automated document classification using machine learning and core network extraction with network analytical measures to identify and remove irrelevant web pages from the network. The article's findings reveal that keyword filtering and automated classification are the most effective techniques for noise reduction, whereas core network extraction did not produce satisfactory results for this particular case. This approach risks excluding other associations related to food safety from the analysis and overlooking important aspects of the actual discourse in the networks. An alternative approach would involve using inductive, data-driven methods such as probabilistic topic models to identify common themes within the text corpus or analyzing word co-occurrences within hyperlink networks. This knowledge could then inform the selection of specific sections of the corpus and network that warrant further in-depth analysis [5].
- A versatile framework for collecting and validating social media data was proposed, applicable beyond Twitter to other public social networking sites and introduced a reporting standard. The inclusion of noise-reducing techniques, such as removing automated content, can be optionally incorporated into search filter development based on the research topic. However, it is not universally necessary for all studies using social media data. The paper emphasizes the importance of transparently disclosing data cleaning and processing methods employed, including whether bot-generated tweets are included or excluded and presents a conceptual framework for collecting social media data, consisting of three key steps: filter development, filter application and filter validation. The article mainly focuses on noise as a distinct methodological challenge in web research during the age of big data. The article also discusses the challenges of estimating precision and recall, especially when accurate human coding and complete data collection are not feasible. However, current trends, future research, or open issues in the field of noise elimination in social media big data are not mentioned in the article [6].
- Sentiment Analysis (SA) was performed on Malay documents by examining three aspects: i) exploring various text representation

methods, ii) emphasizing the impact of pre-processing techniques like normalization and stemming using two types of Malay stemmer algorithms and iii) comparing the performance of Naïve Bayes (NB), Support Vector Machine (SVM) and K-nearest Neighbor (KNN) classifiers in categorizing positive and negative reviews. For sentiment classification, initial steps involve removing incomplete, ambiguous and noisy data through a pre-processing phase that includes tokenization, stop word removal, normalization and stemming. Then, a supervised machine learning classifier is used to categorize sentiments as positive or negative. The results show that the pre-processing strategies slightly improve classifier performance. Their future work aims to develop an auto-translation system to convert dialects into formal Malay, which will be key in easing the normalization process during pre-processing and reducing the time needed to interpret human reviews [7].

In order to assist future researchers in developing innovative algorithms and strategies for noise elimination, we conducted a comprehensive literature survey and analyzed state-of-the-art mechanisms. Thus, the aim of this paper is to review the existing techniques, elucidate their characteristics and outline their strengths and weaknesses. The primary aims of this paper are as follows:

- Examining the current noise elimination mechanisms
- Introducing a novel classification of noise reduction and elimination mechanisms
- Clarifying the merits and drawbacks of noise elimination strategies within each category
- Identifying key areas for further research to enhance noise elimination algorithms.

The following sections of the paper are organized as follows:

Section 2 presents a literature review encompassing the model, metrics, policies and taxonomy pertaining to noise elimination algorithms. Section 3 elucidates the challenges associated with noise elimination algorithms in the context of social media. Section 4 offers an extensive review of existing noise elimination techniques, accompanied by a novel classification framework. Section 5 engages in a discussion regarding the aforementioned techniques, supplemented by relevant statistics. Section 6

outlines open issues that require further examination. Lastly, in Section 7, we conclude our survey and propose potential avenues for future research.

2. Basic Noise Elimination Models, Metrics, and Policies in the Literature

Let us draw a basic diagram illustrating the process of noise elimination in social media big data along with an explanation of each component (Figure 1):

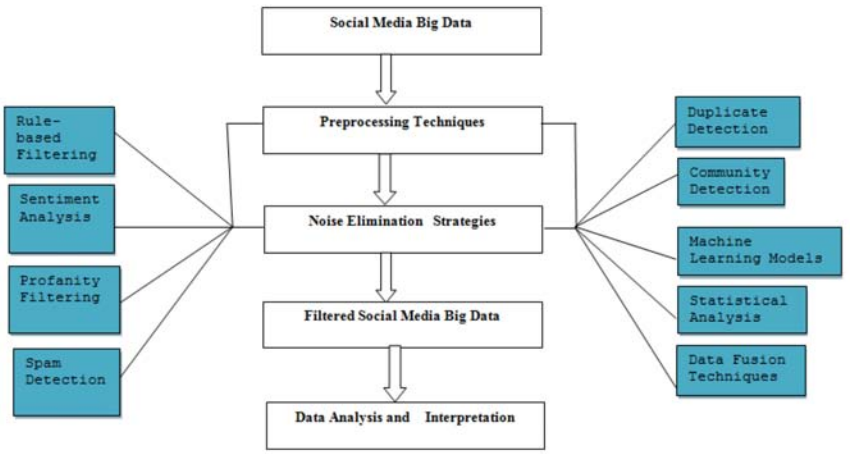

Figure 1

Some Strategies for Noise Elimination in Social Media Big Data

Explanation:

a. **Social Media Big Data**: This is the raw data collected from various social media platforms, including text, images, videos and other multimedia content.

b. **Preprocessing Techniques:** These are ways to clean and prepare the raw data for further analysis. Data cleaning, normalization, tokenization and stemming are examples of preprocessing methods.

c. **Noise Elimination Strategies:** These are algorithms and techniques used to identify and eliminate noise from large amounts of social media data. Rule-based filtering, sentiment analysis, profanity filtering, spam detection and duplicate detection are a few examples of noise removal techniques, which are depicted below:

i. **Rule-based Filtering:** This part uses preset patterns or rules to eliminate noise from social media data. Criteria for detecting spam, foul language, duplicate content, or irrelevant information may be part of these guidelines.

ii. **Sentiment Analysis:** This part examines the feelings or sentiments conveyed in social media posts. When it comes to noise elimination, it can be helpful in determining whether a sentiment is positive, negative, or neutral.

iii. **Profanity Filtering**: This component focuses on identifying and removing offensive or inappropriate language derived from data on social platforms.

iv. **Spam Detection:** This part finds and eliminates spam or unrelated material from social media information. It employs a number of methods, including machine learning models, content analysis and user behaviour analysis.

v. **Duplicate Detection:** Using information from social media platforms, this component finds and eliminates redundant or duplicate content. It helps to improve the quality of the dataset and reduce interference.

vi. **Community Detection:** This feature locates groups or communities on social media platforms. By examining user interactions and connections, it can assist in detecting noise.

vii. **Preprocessing Techniques**: This component includes various preprocessing techniques such as data cleaning, normalization, tokenization and stemming. These methods are utilized to prepare the data for subsequent analysis.

viii. **Machine Learning Models**: This component includes various machine learning algorithms and models used for noise elimination. These models are trained on labeled data to classify or filter out noise from social media data.

ix. **Statistical Analysis**: This component involves statistical techniques and methods used to analyze social media data and identify patterns or anomalies indicative of noise.

x. **Data Fusion Techniques**: This component integrates data from various sources or modalities to improve noise elimination. It integrates data from text, images, videos and other multimedia content to enhance the accuracy of noise detection and removal.

d. **Filtered Social Media Big Data**: This is the data obtained after applying noise elimination strategies. It consists of the cleaned and refined social media data ready for analysis.
e. **Data Analysis and Interpretation**: This stage involves analyzing the filtered social media big data to extract meaningful insights, trends, patterns, or sentiments. Data analysis techniques such as machine learning, Natural Language Processing (NLP) and statistical analysis may be used for this purpose. The interpreted results can then be used for decision-making or further research purposes.

2.1 *Noise Elimination Metrics*

In this subsection, we review the metrics for noise elimination in social media big data. As mentioned before, researchers have proposed several noise reduction algorithms and mechanisms. The literature on noise elimination proposes metrics for applying noise elimination algorithms and we summarize them as follows [39], [40], [41]:

- Accuracy: The proportion of correctly classified instances, measuring the effectiveness of the noise elimination technique in correctly identifying and removing noise.
- Precision and Recall: Precision measures the proportion of correctly identified noise instances among all instances labeled as noise, while recall measures the proportion of correctly identified noise instances among all actual noise instances. These metrics provide insights into the effectiveness and completeness of noise elimination.
- F1-Score: The harmonic mean of precision and recall, which provides a balanced measure of the overall performance of the noise elimination technique.
- RMSE and MAE: RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are common measures used to evaluate the performance of regression models by quantifying prediction errors. RMSE calculates the square root of the average of squared differences between predicted and actual values. It is found by taking the square root of the mean of the squared errors. RMSE is valuable because it emphasizes larger errors due to the squaring process. It offers an overall assessment of the prediction errors, while MAE computes the average of the absolute differences between predicted and actual values, providing a general measure of prediction error magnitude.

2.2 *Taxonomy of Noise Elimination Algorithms and Mechanisms in Social Media Big Data*

In this subsection, we present the existing classification of noise elimination algorithms. In the vast landscape of social media data, various noise elimination algorithms are employed to filter out irrelevant, redundant or misleading information. Figure 2 depicts the Classification of Noise Elimination Strategies in Social Media Big Data. The taxonomy of noise elimination algorithms and mechanisms in social media big data can be classified into several key dimensions as follows:

a. Preprocessing Techniques [9]:

i. Text Cleaning: Methods to remove irrelevant characters, symbols or formatting from social media text data.
ii. Tokenization: Breaking down text into individual tokens (words, phrases, or symbols) for further analysis.
iii. Stop Word Removal: Eliminating common words that do not carry significant meaning in relation to noise elimination.
iv. Stemming and Lemmatization: Reducing words to their root form to handle variations and improve consistency.

Preprocessing methods are vital for readying data for analysis, streamlining its structure and potentially reducing its complexity. However, they may not fully eliminate all forms of noise and might be confined to fundamental text adjustments, possibly leading to the loss of contextual nuances. These techniques find application in various scenarios, such as the initial phases of text manipulation in Natural Language Processing (NLP) endeavors or the preparation of social media content for sentiment analysis or topic modeling.

b. Methods of Filtering [10]:

i. Sentiment analysis: It is the process of determining the emotion or sentiment expressed in social media posts in order to filter out irrelevant content.
ii. Profanity Filtering: Identifying and eliminating offensive or improper language from social media data.
iii. Spam Detection: Recognizing and eliminating spam, including irrelevant or promotional posts.

iv. Duplicate Detection: To cut down on redundancy, duplicate or nearly duplicate posts are found and removed.

v. Bot Detection: Recognizing automated accounts or bots and reducing noise by filtering their activity.

vi. Topic Modeling: This technique uses models such as Latent Dirichlet Allocation (LDA) to identify the primary topics in the data and eliminate irrelevant content.

vii. PageRank-based Filtering: Identifies and eliminates less relevant or poor-quality content by ranking nodes in a graph (such as social media users or posts).

Filtering techniques work well for clearly defined problems, are simple to use and comprehend and are very accurate for some kinds of noise. But they require constant updates to rule sets, require substantial computational resources for large datasets, may not adapt well to new or evolving types of noise, have limited scope in only addressing predefined issues and may misclassify content with ambiguous or mixed sentiments. These techniques can be effective in filtering spam emails or comments on social media platforms, maintaining clean and appropriate user-generated content and removing redundant posts or messages in forums and chat groups, enhancing customer feedback analysis by focusing on relevant sentiments, identifying and removing noise in content analysis tasks.

c. Machine Learning Approaches [11], [12], [37]:

i. Supervised Learning [13]: Training models using labeled data to classify social media posts as noisy or non-noisy.

ii. Unsupervised Learning: Utilizing clustering or anomaly detection algorithms to identify patterns and outliers that represent noise.

iii. Semi-Supervised Learning: Combining labeled and unlabeled data to train models for noise elimination.

These approaches can handle complex and evolving noise patterns, be flexible for accommodating new and unseen data and exhibit high accuracy with sufficient training data. But they require large amounts of labeled data for training (supervised learning) and are computationally intensive. However, the model performance heavily depends on the caliber of training data. In real-time scenarios, these approaches can classify social media posts to detect spam or irrelevant content, identify anomalous user

behavior in social networks, filter noise in user-generated content based on learned patterns.

d. Community Detection [14]:

i. Identifying and analyzing communities or groups of users with similar characteristics or behaviors in social media networks.
ii. Leveraging community structure to identify and filter noise within specific user groups or communities.
iii. Spectral Clustering utilizes eigenvalues of graph-derived matrices to conduct dimensionality reduction prior to clustering in reduced dimensions.
iv. Label Propagation assigns labels based on the labels of neighboring nodes, iteratively updating until a stable state is reached.

Community Detection Algorithms can be effective for network-structured data, reveal hidden patterns and relationships within social networks, and detect outliers or noise based on community structure. But they may not work well for non-network data, are computationally demanding for large networks, their results can be sensitive to the choice of parameters and initial conditions. In practical use, these algorithms can be utilized for detecting fake or spam accounts in social networks, identifying irrelevant or low-quality content in community forums and enhancing social network analysis by filtering out anomalous behavior.

e. Real-Time Monitoring and Feedback [42]:

i. Real-time monitoring of social media data to identify and filter noise as it occurs.
ii. Utilizing user feedback or manual annotation to improve noise elimination algorithms and mechanisms.

f. Evaluation Metrics:

i. Precision, Recall and F1-Score: Assessing the effectiveness of noise elimination algorithms in correctly identifying and removing noise.
ii. User Satisfaction: Considering user feedback and satisfaction with noise elimination results.

g. Statistical Methods:

Outlier Detection: Identifies data points that deviate significantly from the rest of the dataset using statistical measures (e.g., z-scores, Mahalanobis distance). z-scores measure the number of standard deviations a data point is from the mean. Mahalanobis distance considers correlations between variables to measure the distance between a point and a distribution. Simple statistical approaches can be effective for numerical data, identifies extreme deviations that may indicate noise, may not capture complex noise patterns, are limited to specific types of data (e.g., numerical) and can be sensitive to the distribution of data. They detect anomalies in user behavior on social media platforms, identify abnormal posts or interactions that deviate from typical patterns and enhance data quality in Social Media Analytics by removing statistical outliers.

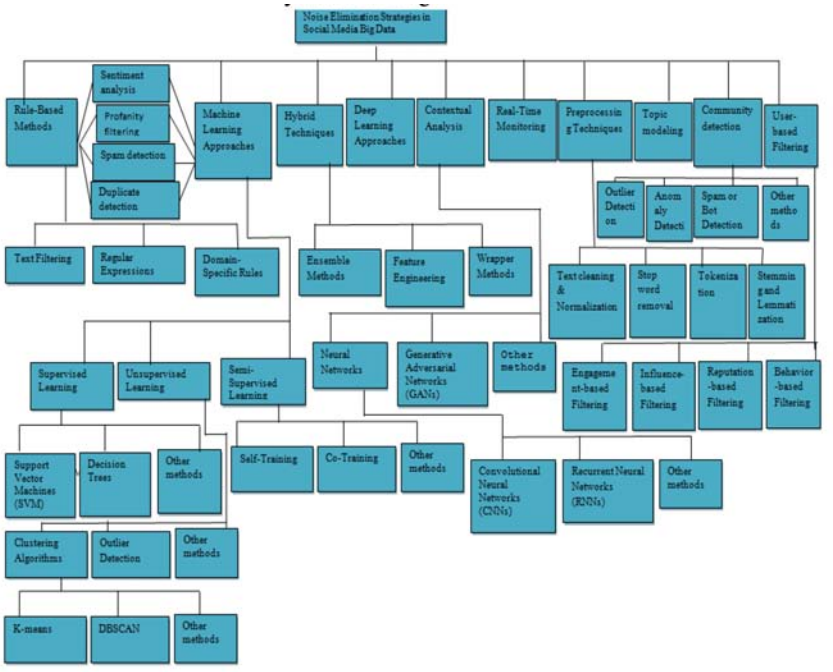

Figure 2

Classification of Noise Elimination Strategies in Social Media Big Data

h. Hybrid Approaches

Hybrid approaches integrate the simplicity and specificity of rule-based methods with the adaptability and accuracy of machine learning techniques, thus improve robustness of noise detection and handle a broader range of noise types. Hybrid approaches integrate the simplicity and specificity of rule-based methods with the adaptability and accuracy of machine learning techniques, thus improving the robustness of noise detection and handling a broader range of noise types.

By employing these diverse algorithms, social media platforms can effectively enhance the data quality and ensure more accurate analytics and insights.

Below table 1 is a summary of noise elimination policies presented in tabular format, highlighting different strategies, their approaches, advantages, disadvantages and use cases.

Table 1

Summary of Noise Elimination Policies

Noise Elimination Policy	Approach	Advantages	Disadvantages	Use Cases
Rule-Based Filtering	Predefined rules to detect and filter noise (e.g., spam, profanity)	Simple to implement and understand	Limited flexibility, requires constant updates	Social media platforms, forums
Machine Learning Filtering	Trained models to identify noise patterns	High accuracy, adaptable to new noise patterns	Requires large datasets and computational resources	News aggregation, brand monitoring
Hybrid Filtering	Combination of rule-based and machine learning techniques	Balances simplicity and accuracy, adaptable	Increased complexity, needs careful integration	Customer service, event monitoring
User Feedback Mechanisms	Collecting user feedback to refine filters	Leverages human judgment for accuracy	Dependence on user participation, potential biases	Social media moderation, community forums

Contd...

Real-Time Monitoring and Feedback	Continuous monitoring with dynamic feedback adjustment	Immediate action, adaptive learning	Resource intensive, complex to implement	Real-time event monitoring, social media platforms
Content-Based Filtering	Analyzing content characteristics (e.g., keywords, context)	Effective for specific types of noise (e.g., fake news)	Can miss nuanced noise, context dependency	News feeds, content recommendation systems
Collaborative Filtering	Using collective input from multiple users to identify noise	Leverages crowd intelligence, scalable	Vulnerable to groupthink, slower response	E-commerce reviews, social media ratings
Spam Detection Algorithms	Specific algorithms to identify and remove spam	Efficient for large volumes of repetitive noise	May need frequent updates, can produce false positives	Email filtering, social media comment sections
Duplicate Detection	Identifying and removing duplicate content	Reduces redundancy, improves data quality	Can be computationally intensive, complex logic required	Data aggregation platforms, content management systems
Profanity Filtering	Detecting and removing offensive language	Maintains platform decency, straightforward implementation	Contextual errors, can be bypassed by users	Social media platforms, online gaming communities
Sentiment Analysis	Analyzing sentiment to filter negative or irrelevant content	Provides insights into user sentiment, versatile applications	Sentiment detection errors, requires contextual understanding	Customer feedback analysis, brand monitoring

3. Challenges in Social Media-Based Noise Elimination Strategies

Review of the literature shows that noise elimination in big data from social media has faced some challenges. Although noise elimination in social media big data has been widely studied, the current state is still far from ideal. In this section, we review the challenges in noise elimination

with the aim of designing improved noise elimination strategies in the future. Some studies have mentioned challenges for the social media-based noise elimination mechanisms ([5], [6], [11], [12], [13], [15], [16], [17], [18], [19]).

Eliminating noise in social media data presents several unique challenges due to the dynamic and unstructured nature of the data. Outlined here are some of the main challenges:

a. Volume, Velocity and Variety of Data:
   - Volume: Social media generates an enormous amount of data every second. Managing and processing such large volumes in real-time to eliminate noise is a significant challenge.
   - Velocity: The rate of new data generation and processing requirements is exceptionally fast. Real-time noise elimination must keep up with this rapid data flow.
   - Variety: Data manifests in diverse formats, such as text, images, videos and audio. Each type requires different noise elimination techniques, adding complexity to the process.

b. Unstructured and Semi-Structured Data: Most of the data from social media is unstructured or semi-structured, making it difficult to apply traditional noise elimination methods that work well on structured data.

c. High Dimensionality: Social media data often includes a large number of features (e.g., hashtags, mentions, multimedia content). Handling high-dimensional data without losing important information while eliminating noise is complex.

d. Dynamic Content and Evolving Language: Language and trends on social media change rapidly. New slang, hashtags and memes emerge frequently. Noise elimination algorithms must continuously adapt to these changes to remain effective.

e. Context Sensitivity: The context of used words or phrases can change their meaning significantly. Algorithms need to understand the context to differentiate between noise and relevant information accurately.

f. Spam and Malicious Content: Identifying and filtering out spam and other forms of malicious content (e.g., phishing links and fake

news) poses a challenge because of the advanced methods employed by malicious actors to avoid detection.

g. Ambiguity and Sarcasm: Social media users often use ambiguous language, sarcasm, or irony, which can confuse noise elimination algorithms and lead to misclassification of relevant content as noise.

h. Privacy Concerns: Analyzing and processing social media data for noise elimination must be done in a way that respects user privacy and complies with regulations like GDPR. Finding the right balance between noise elimination and privacy poses a major challenge.

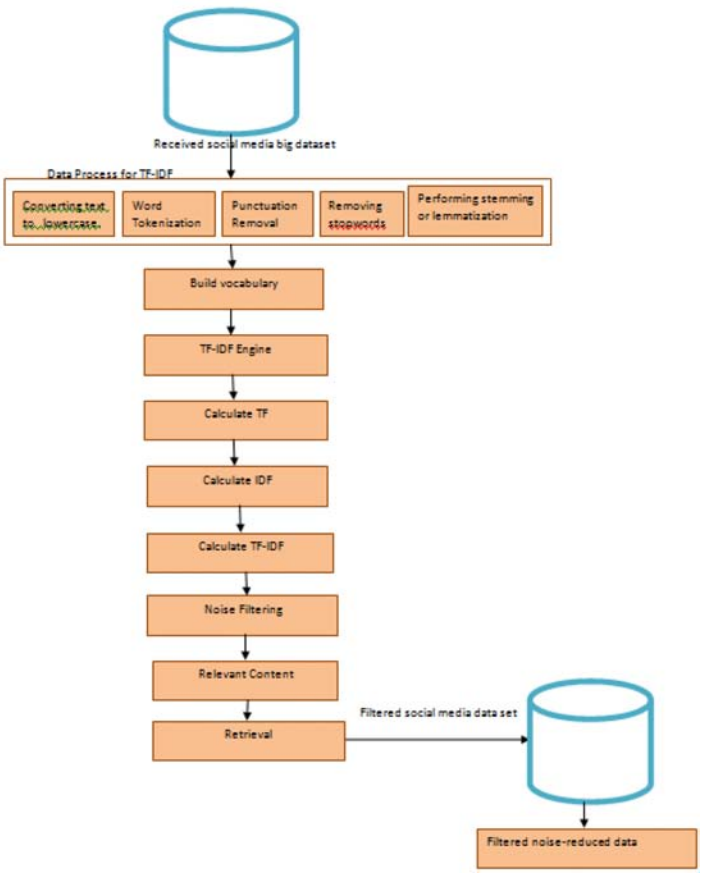

Figure 3

(TF-IDF) based Model for Noise Elimination

i. Multilingual and Multicultural Data: Social media platforms host users from all over the world, leading to content in multiple languages and cultural contexts. Developing noise elimination techniques that work across languages and cultures is complex.
j. Integration of Different Noise Elimination Techniques: Combining various noise elimination methods (e.g., rule-based, machine learning, hybrid approaches) into a coherent system that works effectively across different types of noise and content formats is challenging.
k. Computational Resources: Real-time noise elimination in large-scale social media data requires substantial computational resources, including powerful processors, large memory capacities and efficient algorithms.
l. User Interaction and Feedback: Incorporating user feedback to improve noise elimination algorithms can be useful but also introduces challenges related to collecting, processing and integrating this feedback in a meaningful way.

To address these challenges, researchers and practitioners often employ a combination of techniques, such as NLP, machine learning, sentiment analysis, network analysis, and anomaly detection. Leveraging advanced computational resources, such as cloud computing and distributed processing, can also help in handling the scale and velocity of social media data. Effective noise elimination in big data from social media requires addressing these challenges through innovative and adaptable strategies. Solutions must be scalable, capable of adjusting to different contexts and equipped to manage the varied and ever-changing content found on social media platforms. Balancing accuracy, efficiency and privacy is crucial to developing robust noise elimination systems that can keep pace with the ever-changing landscape of social media.

4. Survey on existing noise elimination mechanisms in social media big data

Here are some examples of existing noise elimination algorithms and mechanisms in big data from social media with detailed explanation, along with their characteristics, strengths, weaknesses and real-life applications with the help of diagrams where applicable.

4.1 *TF-IDF (Term Frequency-Inverse Document Frequency) ([14], [20], [40]):*

a) Characteristics:
   - TF-IDF assesses the significance of a word in a document set.
   - It is calculated as the product of two metrics: term frequency (TF) and inverse document frequency (IDF).
   - Term frequency (TF) gauges how often a word appears in a document, while inverse document frequency (IDF) measures the importance of the word across the entire collection of documents.

     tf-idf(t,d) = tf(t, d) * idf(t) where:

     tf(t, d) = Number of times term t appears in document d

     idf(t)=log(Total number of documents / Number of documents with term t in it)

b) Advantages: Effective in identifying and removing irrelevant or low-quality content, such as spam and low-information posts.

c) Disadvantages: Might not encompass contextual nuances and could be susceptible to variations or changes in vocabulary over time.

d) Real-life application: Applied in social media surveillance and content curation to filter out noise and focus on the most relevant and informative posts.

TF-IDF can be implemented in four steps: 1: Data Pre-processing, 2: Calculating Term Frequency, 3: Calculating Inverse Document Frequency, 4: Calculating Product of Term Frequency & Inverse Document Frequency. (TF-IDF) based Model for noise elimination is shown in Figure 3

4.2 *Latent Dirichlet Allocation (LDA) ([19], [21]):*

a) Characteristics: LDA is a topic modeling algorithm that identifies latent topics in a collection of documents.
   - It assumes that each document is a mixture of various topics and each topic is a distribution over a fixed vocabulary of terms.
   - The algorithm learns the topic distributions and the topic proportions for each document in an unsupervised manner.

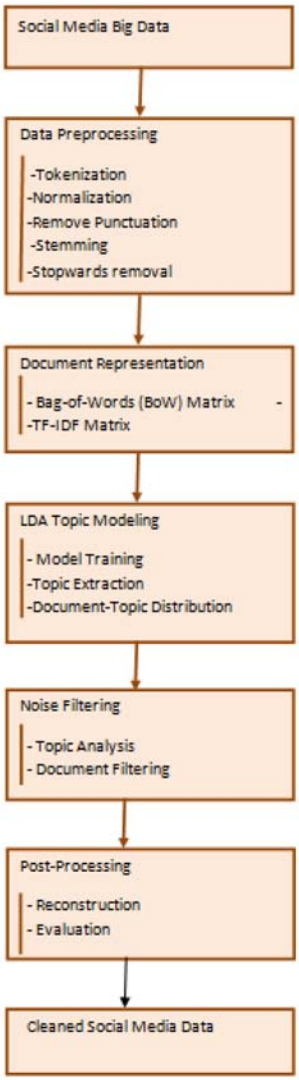

Figure 4

The Role of LDA in Noise Elimination from Social Media Big Data

b) Advantages:

i) Effective in detecting and removing content that does not align with the dominant topics, such as off-topic or irrelevant posts.

ii) Useful for identifying and grouping similar content, enabling better filtering and curation.

c) Disadvantages:

i) It is computationally intensive and may require careful parameter tuning.

ii) May not capture all the nuances and context of the content, leading to potential misclassification.

d) Real-life application: Used in social media analytics to identify and remove noise by grouping and filtering content based on its relevance to the topic.

The diagram (Figure 4) illustrates the role of LDA in noise elimination from social media big data:

Social Media Big Data: The unprocessed data sourced directly from social media platforms, including noisy and irrelevant information.

1. Data Preprocessing: Initial steps to clean and prepare the data for analysis, removing extraneous elements.
2. Document Representation: Conversion of text data into numerical formats suitable for topic modeling.
3. LDA Topic Modeling: Application of LDA to discover topics and represent documents as distributions over these topics.
4. Noise Filtering: Using the topic distributions to filter out documents associated with irrelevant or noisy topics.
5. Post-Processing: Reconstructing the dataset with filtered data and assessing the effectiveness of the noise reduction process.
6. Cleaned Social Media Data: The final output consisting of social media data with reduced noise, focused on relevant topics.

By leveraging LDA in this structured approach, we can notably improve the caliber of data in social media, making it more useful for subsequent analysis and decision-making processes.

4.3. *Anomaly Detection using One-Class Support Vector Machines (OC-SVM) [43]*

a) Characteristics:

i) Anomaly detection using One-Class Support Vector Machines (OC-SVM) is a potent method for identifying and eliminating noise in social media big data.

ii) The OC-SVM is trained on a dataset comprising only 'normal' data, allowing it to detect and filter out anomalous data points that exhibit significant deviation from the norm.

b) Advantages: Effective in detecting and removing spam, bots, irrelevant content, fake accounts, or other anomalous activities on social media platforms.

c) Disadvantages: May require a well-defined set of normal data to train the model effectively.

d) Real-life application: Applied in the detection of fraudulent activities and moderation of content in social media to identify and remove harmful or suspicious content.

e) Process and Components:

1. Data Collection: Collect extensive amounts of social media content, such as posts, comments, and user interactions.
2. Data Preprocessing:
   i. Text Cleaning: Remove unnecessary elements, such as stop words, punctuation and special characters.
   ii. Tokenization: Split text into individual tokens or words.
   iii. Feature Extraction: Convert text data into numerical features suitable for model training (e.g., TF-IDF, word embeddings).
3. Training Data Preparation:
   i. Select a subset of data representing regular or relevant information.
   ii. This subset functions as the training dataset for the OC-SVM.
4. Model Training (OC-SVM): Train the OC-SVM model on the normal data to learn its characteristics and boundaries.
5. Anomaly Detection:
   i. Apply the trained OC-SVM model to the entire social media dataset.
   ii. The model classifies each data point as regular or anomalous.
6. Noise Filtering: Remove the data points identified as anomalies (noise) by the OC-SVM.
7. Post-Processing:
   i. Reconstruction: Compile the cleaned dataset containing only normal data points.

ii. Evaluation: Assess the quality and relevance of the cleaned data.

Figure 5 illustrates the role of OC-SVM in noise elimination from social media big data.

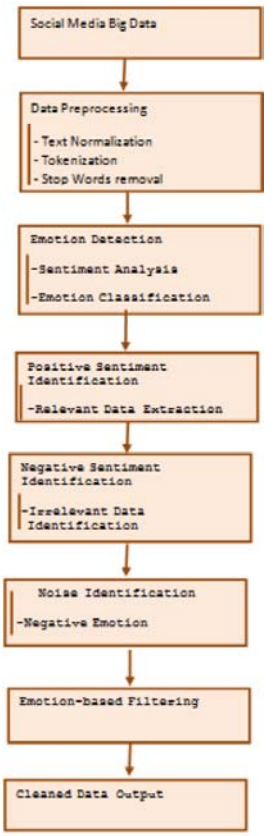

Figure 5

The Role of OC-SVM in Noise Elimination from Social Media Big Data

4.4 *Emotion-based Filtering ([22], [23], [24])*

a) Characteristics:

1. Emotion-based Filtering is a method employed to enhance the caliber of social media data by identifying and filtering out noisy data based on emotional content.

2. The process involves analyzing the emotional tone of user-generated content, such as comments, posts and reviews and removing or flagging content that does not align with the desired emotional spectrum. This approach helps in maintaining a consistent quality of data, which is essential for various analytical objectives.

b) Advantages: Filtering out noisy data improves dataset quality and relevance, enables focused analysis for accurate insights, automates processing of large data volumes, and effectively removes non-informative content like trolling or sarcasm.

c) Disadvantages: Implementing effective emotion detection and filtering algorithms is complex, potentially biased, computationally intensive and difficult because of the informal and varied language used in social media.

d) Real-life application: Emotion-based filtering in monitoring of social media and customer service helps identify relevant customer feedback, enhances brand monitoring by focusing on constructive comments and aids mental health analysis by filtering data to meet specific emotional criteria.

Figure 6 demonstrates the role of Emotion-based Filtering in noise elimination from social media big data:

4.4.1. Components of Emotion-based Filtering

a. Social Media Data: The unprocessed data gathered from diverse social media platforms, encompassing comments, posts, reviews, tweets, etc.

b. Data Preprocessing: Initial cleaning steps such as removing duplicates, handling missing values and normalizing text to prepare the data for further analysis.

c. Emotion Detection: This component involves:

   i) Sentiment Analysis: Determining the sentiment polarity (positive, negative, neutral) of the text.

   ii) Emotion Classification: Identifying specific emotions conveyed in the text, like joy, sorrow, frustration, fears, etc.

d. Positive Sentiment Identification: Extracting data with positive sentiments, this is often considered relevant and high-quality.

e. Negative Sentiment Identification: Identifying data with negative sentiments, which may include noise or irrelevant content.

f. Noise Identification: Using the results from the emotion detection step, this component identifies noisy data. Noisy data in this context could be:

   i) Negative Emotion: that is overly negative or toxic, which might not be useful for the intended analysis.

   ii) Irrelevant Emotion: Content with emotions which are irrelevant to the analysis goals.

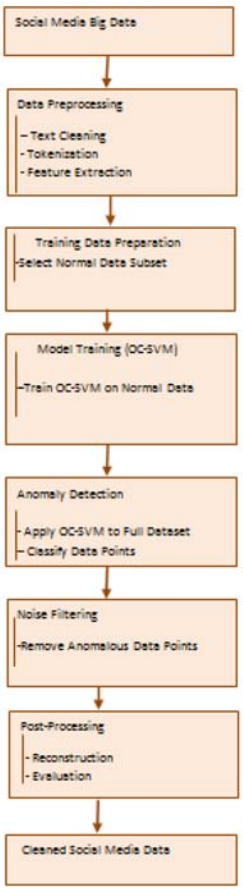

Figure 6

The Role of Emotion-based Filtering in Noise Elimination from Social Media Big Data

g. Emotion-based Filtering: The identified noisy data is filtered out. This could involve removing the data or flagging it for further review, based on the particular needs of the analysis.
h. Cleaned Data Output: The resultant dataset after filtering, which is anticipated to exhibit superior quality and be more relevant for analysis.

Emotion-based Filtering is a powerful tool for enhancing the quality of social media data, rendering it more suitable for various analytical purposes by removing emotionally irrelevant or noisy data.

4.5. *Graph-based Approaches* ([25], [26]):

a) Characteristics:
   i. Relational Modeling: Graph-based models represent social media data as networks, capturing complex relationships and interactions, which helps identify patterns and anomalies not evident in flat data models.
   ii. Scalability and Unsupervised Learning: These models handle massive social media data scales using distributed processing and can identify patterns and outliers without labeled data, beneficial in evolving noise characteristics.
   iii. Contextual Understanding and Temporal Dynamics: They incorporate semantic context, capturing evolving data dynamics to filter context-dependent noise and are robust to noise and missing data, providing interpretable insights and multidimensional analysis. To illustrate the role of graph-based approaches in noise elimination from social media big data, we can create a flow diagram that visually represents the process and explains the different components as follows:
   1. Social Media Data Collection: Aggregating raw data from multiple social media platforms, including posts, comments and interactions.
   2. Data Preprocessing: Preparing the data for analysis. This involves multiple steps: i) Text Normalization: Standardizing text (lowercasing, removing punctuation, etc.). ii) Tokenization: Segmenting text into individual words or phrases. iii) Stop Words Removal: Eliminating common, non-informative words (e.g., “the”, “and”, “is”).

3. Graph Construction: Creating a graph structure where nodes represent data points (e.g., users, posts) and edges represent relationships or interactions.
4. Community Detection: Identifying clusters within the graph to group similar data points, analyze community properties for anomalies and treat unusual communities as potential sources of noise in social media data.
5. Anomaly Detection: Using graph-based algorithms to identify outliers within communities, associate anomalies with potential noise and prioritize their analysis and elimination to enhance the overall noise detection and filtering process.
6. Label Propagation: Propagating known labels through the graph to infer labels for unlabeled nodes, using algorithms like Random Walk with Restart, Harmonic Function and Label Spreading, helps identify potential noise based on the spread and confidence of the propagated labels.
7. Noise Identification: Isolating data points identified as noise based on their anomalous behavior or inconsistency with community labels, using structural, temporal and behavioral analysis, to prioritize and eliminate the most problematic elements in the social media graph.
8. Graph-based Filtering: Applying filtering techniques to remove or correct noisy data points identified in the previous step, using noise scoring or ranking mechanisms based on graph-based features and leveraging anomaly detection, community detection and centrality measures to selectively filter high-noise elements while preserving the overall graph structure.

   Advantages of Graph-based Approaches in Noise Elimination from Big Data in Social Media: Graph-based techniques can effectively detect and remove content from suspicious or malicious accounts, as well as identify and eliminate low-quality connections by leveraging structural insights, community detection, scalability, rich representations, flexibility and anomaly detection.

   Disadvantages of Graph-based Approaches in Noise Elimination from Big Data in Social Media: Graph-based noise elimination methods require extensive data on the social network structure, which might not always be accessible and face challenges such

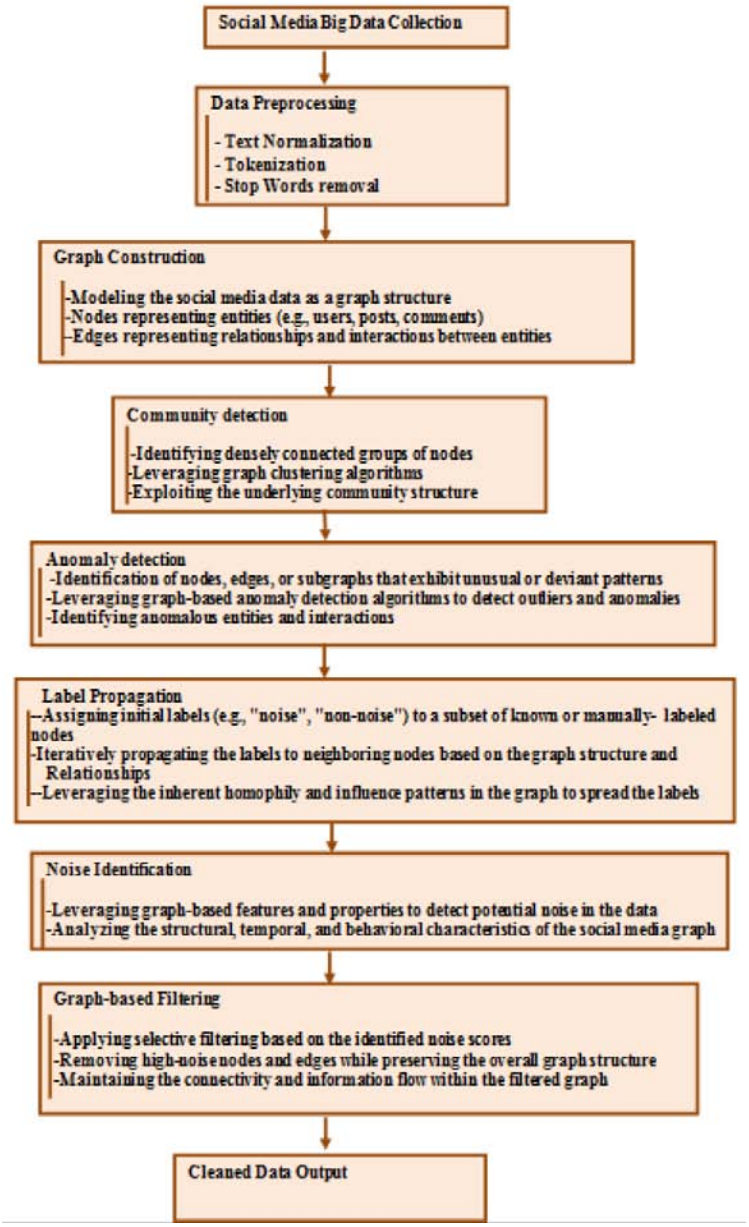

Figure 7

The Role of Graph-based Approaches in Noise Elimination from Social Media Big Data

as complexity, overfitting, data quality dependence, algorithm sensitivity, interpretability issues and resource intensity, impacting real-life applications like social media platform moderation.

Example Use Case: For Twitter spam detection, leveraging graph-based methods offers advantages like gaining structural insights and detecting spam communities, but faces challenges in complexity and dependence on data quality, given the massive scale of interactions.

10. Clean Data Output: Producing a refined dataset with reduced noise, ready for further analysis or modeling.

This diagram (Figure 7) and the accompanying explanation provide a clear overview of how graph-based approaches can effectively eliminate noise from social media big data. Each step in the process contributes to improving the data quality, making it more reliable for subsequent analysis.

The following Table 2 provides a brief overview of past studies on social media big data-based noise elimination techniques:

Table 2
An Overview of Social Media Big Data-based Noise Elimination Techniques

Year	Author (s)	Key Ideas	Objectives	Advantages	Disadvantages	Evaluation	Journal/ Conference
2024	Lia & Zhu	Lit. review, noise taxonomy, algo dev, benchmark	Review DL for noisy labels; create synthetic benchmark	Systematic, real-world, benchmark	Complex, narrow focus, limited empirics	Test noise-robust methods on benchmark	Elsevier, arXiv
2024	Pampapura Madali et al.	Data collect, sentiment, topic model, processing	Study misinformation & social noise in BLM	Rich data, public insight	Limited scope, no multimedia	Sentiment, topic modelling	Journal of Information Science
2023	Talaat et al.	DAM, DPM, PRM, ERM, feature select, HPO, multiclass	Personality & emotion recognition from text	High accuracy, full system	Language limit, data issues	Accuracy comparison	Neural Computing & Applications
2023	Toufique Elahi	Annotation, multi-label, noise reduction, sentiment	Test noise reduction on Bangla text	Dataset, noise types	Weak reduction, imbalance	Comparative + human eval	Noisy User-generated Text Workshop
2023	Jain & Xu	AIWrap, AI feature select, real/sim tests	Improve feature selection in high-dimensional data	Less compute, better features	Overfit risk, complex	Metrics, RMSE compare	BMC Bioinformatics

Contd...

2023	Pau et al.	Preprocess, feature select, noise inject, consistency	Test feature selection under class noise	Deep evaluation	Artificial noise, single classifier	Stability + performance	MDPI Information
2023	Alabduljabbar et al.	Preprocess, feature extract, recommender, web UI	Context-aware news recommendation	Personalization, real use	Small data, short eval	Metrics, user feedback	IJ Computational Intelligence
2023	Lee, Kim & Kim	Graph SSL, KNN graph, noise shaving, label propagation	Improve graph semi-supervised learning	Better accuracy, less tuning	High compute, complex	Tests on real/sim data	Applied Soft Computing
2023	Mahl et al.	Lit. review, search term validation, topic analysis	Validate search terms in journalism data	Reproducible, robust data	Noise from bad terms	Precision, recall, F1	Digital Journalism
2022	Prasad Jasti et al.	RBFR, multi FS compare, 5 ML models	Rank relevant text features	+25% accuracy, robust FS	Threshold tuning, 3 datasets	Acc, P, R, F1, 10-fold	Journal of Nanomaterials
2022	Johnson & Khoshgoftaar	Survey label noise, DL, streaming, big data	Survey class noise handling in big data	Broad, practical, open source	Very broad, complex	Empirical method compare	J. Data & Information Quality
2022	Brambilla et al.	Keyword + ML, HITL, graph model, LDA, stance	Analyze social media discussions	Multi-analysis view	Complex, human effort	Accuracy on real data	Big Data Cognitive Computing
2021	Mayiami et al.	GMRF, graph filter, optimization, proximal method	Infer graph from noisy signals	Accurate, efficient	Complex, assumptions	NMSD, NMI, F-measure	Discover IoT

Contd...

2021	Jane & Arockiam	DaRoN, central tendency, SVM	Remove noise in IoT agriculture data	Better quality, automation	Residual noise, complex	Accuracy compare	Annals Romanian Soc. Cell Biology
2021	Azzahra et al.	SVM, TF-IDF, Chi-square, preprocessing	Toxic comment classification	Automates moderation	Imbalance, time	F1-score	IJICT
2021	Caiafa et al.	ANN, DA, NMF, EMD, FS, DL, ethnography	Handle imperfect datasets in ML	Practical cases, new methods	Heavy compute, less general	Real dataset validation	Applied Sciences
2021	Zimmerman	Social noise concept, ethnography, FB observation, interviews, thematic analysis	Study surveillance effect on info-seeking & social capital	Explains behavior change, theory model	Limited geography, no demographics	4 social noise components analysis	Journal of Documentation
2021	Sharou Li & Specia	NLP noise taxonomy, sources, preprocessing impact	Promote task-specific handling of non-standard text	Language-independent taxonomy, tailored approach	Limited tasks, partial generalization	Experiments vs standard preprocessing	RANLP 2021
2021	Al-Gethami et al.	ML (DT, RF, SVM, ANN, NB), ensembles, noise filtering	Study noise effect on IDS accuracy	Tests ML robustness, filtering study	Limited datasets, basic filters	Accuracy: split, test set, 10-fold	Security & Communication Networks
2020	Oleghe	Sensor data, ReliefF FS, RF prediction, WEKA, Python	Replace faulty sensor values	Preserves data, better quality	High compute, case-specific	Prediction accuracy	Journal of Big Data

Contd...

2020	Nematzadeh et al.	K-means, filtering algos, remove/relabel noise	Propose CLCF for class noise	Accuracy gain after filtering	Binary only, weak outlier handling	Accuracy on 4 datasets	Springer Applied Sciences
2019	Saseendran et al.	Regression, RF classifier, noise impact study	Analyze noise on ML models	Shows algorithm resilience	Only accuracy studied	RMSE, Accuracy	CC BY-NC-SA
2019	Gupta & Gupta	Ensemble, single, polishing, filtering, robust methods	Review noise handling methods	Ensemble best accuracy	Polishing may add errors	Comparative study	ISICO 2019
2018	Jittawiriyanukoo	Regression, DEL, SAM, RAM for noise/missing	Study filtering in big data	Simple, low bias tools	Risk of data loss	RMSE, COEF, MOA compare	IJECE
2018	Baharloua & Aghamaleki	CNN TL, 6 DL models, Isolation Forest	Vehicle recognition with noisy web data	No large labels needed, better accuracy	Limited scope, small data	Accuracy before/after cleaning	ESWA
2018	Onyancha & Plekhanova	ML for dynamic web noise, user interest modeling	Reduce web noise via user behavior	Improves user profile quality	No detailed results	Compare with existing methods	WASET IJ
2017	Ekambaram	SVM SVs, OCSVM, TCSVM, subset select	Detect/remove label noise	Captures most noise, less review	Misses some, manual step	Tests incl. ImageNet	USF Tampa Thesis
2017	Onyancha & Plekhanova	ML for dynamic web noise, user interest	Improve web profile by smart noise removal	Adapts to user change	Needs logs, high computational cost	Compare with existing	Zenodo (CERN)

Contd...

2016	Arif & Mustapha	Text representation normalization, Malay stemming, NB/SVM/kNN	Improve Malay SA via preprocessing	Shows preprocessing impact	Single dataset, no DL	Classifier comparison	iCMS Papers
2016	Kaur & Maini	Linear/nonlinear filters, median, adaptive	Noise removal in cell/digital images	Filter comparison, quality gain	40 images, few noises	MSE, PSNR, CoC, visual	IJIGSP
2016	Waldherr et al.	Keyword filter, ML classify, network analysis	Reduce noise in hyperlink issue networks	Keyword & ML effective	15% less than human, narrow scope	Empirical filter compare	Social Science Computer Review
2016	Li et al.	Crowd labels, MV/RY/ZenCrowd, noise filters	Improve crowdsourcing labels	Lowers label noise	Residual noise	14 UCI + 3 real sets	Knowledge-Based Systems
2016	Uddin & Lee	NR-and-SDD, probabilistic model	Clean unstructured social data for student prediction	Better structured data	Early stage, noise issues	NR/SDD efficiency	ASONAM (IEEE/ACM)
2015	Bao, Li & Xiong	SLRA, Cadzow, Hankel preserve	Remove noise for modal analysis	Better damping estimates	Needs threshold choice	Modal param compare	Applied Physics Letters
2015	Sáez et al.	Ensemble & iterative filtering, noise score	Improve class accuracy after filtering	Balance clean/noisy removal	Threshold sensitivity	Compare with filters	Information Fusion
2014	Frenay & Verleysen	Survey label-noise methods	Taxonomy & robust algorithms	Practical guidance	Full robustness rare	Misclassification prob.	IEEE TNNLS

Contd...

2014	Wang et al.	NTCD, dynamic communities	Detect evolving communities, anomalies	Uses past and current data	Needs real validation	Compare with OSLOM	J. Combinatorial, Optimization
2014	Brynielsson et al.	SVM, NB, rules, tweet emotion	Crisis tweet emotion classification	Real-time tool, good SVM	Low tweet reliability	Classifier accuracy	Security Informatics
2013	Baldwin et al.	Linguistic stats, parser, OOV	Measure social media noisiness	NLP insights	Corpus bias	cross-platform comparison	IJCNLP
2013	Narayana et al.	Negation, n-grams, MI, NB	Improve NB sentiment accuracy	88.8% IMDB accuracy	Few limits stated	Accuracy measure	IDEA Conf.
2013	Natarajan et al.	Loss correction, unbiased estimator	Noise-tolerant binary learning	Proven bounds, simple	Only random noise	Benchmarks, synthetic	NeurIPS
2012	Fefilatyev et al.	SVM SV + human check	Detect label noise efficiently	HITL, efficient	SVM-only, bias risk	Detection accuracy	ICPR 2012
2012	Garcia et al.	Noise filter, classifier agreement	Detect/remove class noise	Precise, controlled	High compute, class bias	RENN baseline	SBRN
2006	Xiong et al.	Hyperclique (HCleaner), outlier/ cluster noise handling	Improve clustering & association by noise mitigation	Simple, local density, robust, adaptable	Parameter-sensitive, high compute, binary-only	Compare with distance, clustering, LOF methods	IEEE TKDE
2003	Chen, Huang & Benesty	Adaptive filtering for speech noise	Reduce noise in speech communication	Works for non-stationary noise, single sensor	Needs good signal capture, cancellation risk	Beamforming, ANC, spectral modification	Signals & Communication Technology

([1],[2],[3],[4],[5],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[18],[19],[20],[23],[24],[25],[27],[28], [29],[30],[31],[32],[33],[34],[35],[36],[37],[38],[39],[40],[43])

5. Discussion and Statistics

In this section, we present helpful insights from the articles we've analyzed. Figure 8 effectively illustrates the distribution of these articles across different years, from 2003 to April 2024. The figure shows that each slice represents the number of articles published in that year; for example, the highest counts are in 2016 and 2021, each with 8 articles. The figure also highlights the annual percentage of articles published. Notably, 2018 and 2023 stand out with significant numbers. Specifically, 5% of the articles were published in 2014 and 2015, 14% in 2016, 11% in 2018, another 14% in 2021 and 11% in 2023. Overall, this means that 61% of the articles have been published in the past seven years, showing significant recent growth in research.

The distribution of the studied articles from different publishers is shown in Figure 9. In the figure, the article frequency for each publisher is represented on the corresponding slice, with 13% of journals belonging to IEEE. To further analyze the primary publication sources of the articles, 7% of the literature is associated with Elsevier, 13% with Springer, 4% with ACM, 7% with SAGE Journals, 7% with MDPI open-access scientific journals, 7% indexed on ResearchGate (a popular online hub) and 43% attributed to others.

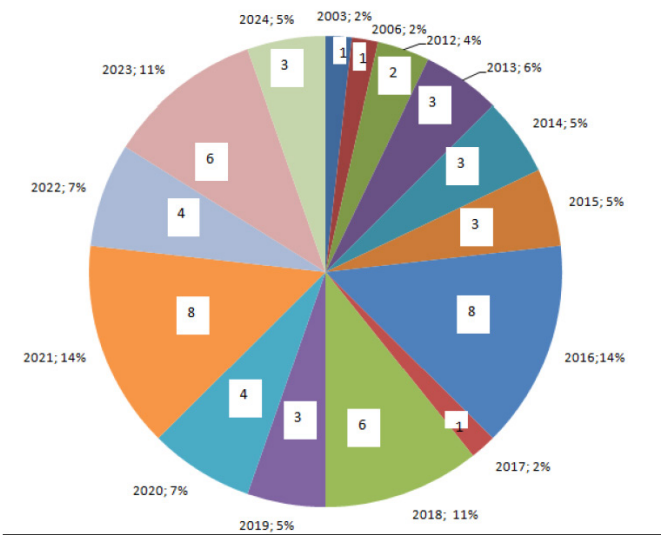

Figure 8

The Distribution of Studied Articles over Time from 2003 until April 2024

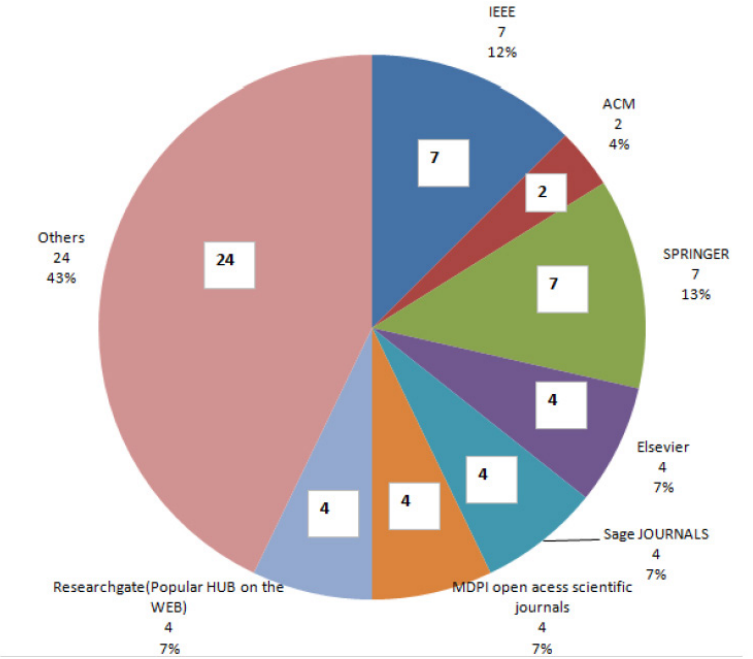

Figure 9

The Distribution of Studied Articles based on Different Publishers

In figure 10, the article frequency for each technique is represented on the corresponding slice,it shows that around 5% of the studied techniques used TF-IDF, 5% used Latent Dirichlet Allocation (LDA), 5% applied anomaly detection using OC-SVM, 5% focused on emotion-based filtering, 5% used graph-based approaches, and 5% applied Natural Language Processing (NLP). Hybrid approaches combining two or more methods accounted for about 31%, while the remaining 39% included other techniques or survey studies related to noise reduction or removal in social media big data. We see that the majority of studies have concentrated on various other techniques like autoencoder, deep learning, label noise-robust methods, Deep Neural Networks, Sentiment Analysis, methods for treating label noise in big data contexts, Multi-step approach for validating search terms, Data exploration, preparation, parsing, feature selection techniques, Measures of Central Tendency, Noise Web Data Learning (NWDL) algorithm**,** different filtering techniques, Noise-Tolerance Community Detection, Structured Low Rank Approximation, Neural Networks (CNNs), various classification algorithms, etc.

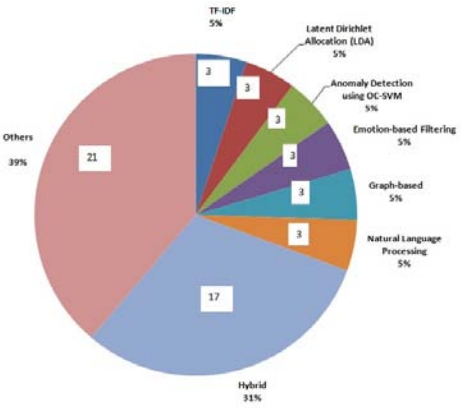

Figure 10

An Overview of Types of Noise Elimination Mechanisms Addressed by the Reviewed Techniques

6. Open Issues in Noise Elimination in Social Media Big Data

Addressing noise elimination in social media big data presents several challenges, including data quality and availability issues such as incomplete datasets and access limitations resulting from privacy policies. Algorithm complexity and scalability are also significant challenges, as many algorithms are computationally intensive and struggle to handle the large-scale, high-velocity nature of social media data. The dynamic and evolving nature of social media necessitates adaptive algorithms that can comprehend context and adapt to evolving semantics. Bias and fairness concerns arise from potential algorithmic bias and the risk of over-filtering, which can result in the removal of legitimate content. The interpretability and transparency of many advanced algorithms, particularly those based on machine learning, are limited, eroding user trust. Integrating multimodal data from various sources adds complexity, as does ensuring real-time processing with low latency. User and content diversity, including language, cultural variations, and behavioral differences, further complicate noise detection. Ethical and privacy concerns must be balanced against the need for noise elimination. Standardized benchmarks and robust validation methods are necessary to evaluate the efficacy of noise elimination techniques. Further research in these areas is crucial for advancing noise elimination in social media big data, ensuring accuracy, efficiency and fairness.

The combined Fuzzy Naïve Bayesian approach, coupled with a supervised FS wrapper-based bidirectional elimination technique, can enhance predictive classification accuracy by selecting essential features and removing meaningless, incorrect and erroneous data. This hybrid method leverages fuzzy logic to handle uncertainty and Naïve Bayes classification for probabilistic predictions, while the supervised FS wrapper iteratively adds and removes features to optimize performance. Our survey reveals that only a few articles have addressed these topics. By refining the dataset and focusing on relevant information, this approach minimizes noise, making it a promising avenue for future research in social media, particularly for achieving significant noise reduction in social media data.

7. Conclusion and Future Work

This survey has comprehensively reviewed various noise elimination techniques applied to social media big data, highlighting their strengths, weaknesses and appropriate use cases. Social media data, characterized by its high volume, velocity and variety, inherently contains significant noise, which can adversely affect data analysis and insights. We categorized noise elimination methods into several major approaches, including machine learning-based algorithms, natural language processing (NLP) techniques, signal processing methods, statistical approaches, graph-based algorithms, heuristic and rule-based methods, hybrid techniques and deep learning approaches. Each category was evaluated in terms of its effectiveness, computational efficiency and relevance to various forms of noise and data structures in social media. Our examination emphasizes the importance of selecting the appropriate noise reduction method, taking into account the unique characteristics of the dataset and the type of noise present.

The evolving landscape of social media and the continuous influx of new data present ongoing challenges and opportunities for research on noise elimination. Future work can focus on several key areas: developing advanced deep learning models like transformers to improve noise detection in textual and multimedia data; enhancing real-time noise elimination techniques for immediate filtering during live events; investigating methods to eliminate noise from data aggregated across multiple platforms; creating context-aware algorithms to filter noise based on data context adaptively; improving scalability and robustness to handle growing data volumes; leveraging user feedback and community-driven

signals for algorithm refinement; and exploring multi-modal noise elimination techniques to process and filter noise across text, images, videos and audio simultaneously.

In conclusion, while significant progress has been made in developing noise elimination techniques for social media big data, the dynamic nature of social media requires ongoing research and innovation. By addressing the outlined future directions, the discipline can further enhance the quality and reliability of insights extracted from social media, ultimately supporting better decision-making and a deeper understanding of social phenomena.

References

[1] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing big social data," *J. Big Data,* vol. 4, Art. no. 3 (Jan. 2017), doi: 10.1186/s40537-017-0063-x.

[2] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing data analysis with noise removal," *IEEE Trans. Knowl. Data Eng*., vol. 18, no. 3, pp. 304–319 (Mar. 2006), doi: 10.1109/TKDE.2006.46.

[3] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naïve Bayes model," in Proc. Int. Conf. Intelligent Data Engineering and Automated Learning (IDEAL), Hefei, China, pp. 194–201 (2013), doi: 10.1007/978-3-642-41278-3_24.

[4] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, "INFFC: An iterative class noise filter based on the fusion of classifiers," Inf. Fusion, vol. 27, pp. 19–32 (Jan. 2016), doi: 10.1016/j.inffus.2015.04.002.

[5] A. Waldherr, D. Maier, P. Miltner, and E. Günther, "Big data, big noise: The challenge of finding issue networks on the web," Soc. Sci. Comput. Rev., vol. 35, no. 4, pp. 427–443 (Aug. 2017), doi: 10.1177/0894439316643050.

[6] Y. Kim, J. Huang, and S. Emery, "Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data," *J. Med. Internet Res.*, vol. 18, no. 2, Art. no. e41 (Feb. 2016), doi: 10.2196/jmir.4738.

[7] S. M. Arif and M. Mustapha, "The effect of noise elimination and stemming in sentiment analysis for Malay documents," in Adv. Intell. Syst. Comput., Singapore, pp. 115–124 (2016), doi: 10.1007/978-981-10-2772-7_10.

[8] R. Kaur and R. Maini, "Performance evaluation and comparative analysis of different filters for noise reduction," *Int. J. Image Graph. Signal Process.*, vol. 8, no. 7, pp. 9–21 (Jul. 2016), doi: 10.5815/ijigsp.2016.07.02.

[9] K. Al Sharou, Z. Li, and L. Specia, "Towards a better understanding of noise in natural language processing," in Proc. Recent Advances in Natural Language Processing (RANLP), Online, pp. 53–62 (Sep. 2021), doi: 10.26615/978-954-452-072-4_007.

[10] J. Chen, Y. Huang, and J. Benesty, "Filtering techniques for noise reduction and speech enhancement," in Adaptive Signal Processing: Applications to Real-World Problems, Berlin, Germany: Springer, pp. 129–154 (2003), doi: 10.1007/978-3-662-11028-7_5.

[11] Cesar F. Caiafa , Zhe Sun ,Toshihisa Tanaka , Pere Marti-Puig and Jordi Solé-Casals, "Machine learning methods with noisy, incomplete or small datasets," *Appl. Sci.*, vol. 11, no. 9, Art. no. 4132 (May 2021), doi: 10.3390/app11094132.

[12] S. Gupta and A. Gupta, "Dealing with noise problem in machine learning datasets: A systematic review," *Procedia Comput. Sci.*, vol. 161, pp. 466–474 (Nov. 2019), doi: 10.1016/j.procs.2019.11.146.

[13] Khalid M. Al-Gethami, Mousa T. Al-Akhras, and Mohammed Alawairdhi, "Empirical Evaluation of Noise Influence on Supervised Machine Learning Algorithms Using Intrusion Detection Datasets," *Secur. Commun. Netw.*, vol. 2021, Art. no. 8836057 (2021), doi: 10.1155/2021/8836057.

[14] Li Wang, Jiang Wang, Yuanjun Bi, Weili Wu, Wen Xu, and Biao Lian, "Noise-tolerance community detection and evolution in dynamic social networks," *J. Comb. Optim.*, vol. 28, no. 3, pp. 600–612 (May 2014), doi: 10.1007/s10878-014-9719-z.

[15] O. Oleghe, "A predictive noise correction methodology for manufacturing process datasets," *J. Big Data*, vol. 7, Art. no. 67 (2020), doi: 10.1186/s40537-020-00367-w.

[16] Daniela Mahl, Gerret von Nordheim, and Lars Guenther, "Noise Pollution: A Multi-Step Approach to Assessing the Consequences of (Not) Validating Search Terms on Automated Content Analyses," *Digital Journalism*, vol. 11, no. 5, pp. 1–23 (2023), doi: 10.1080/21670811.2022.2114920.

[17] Diego García-Gil, Julián Luengo, Salvador García, and Francisco Herrera, "Enabling smart data: Noise filtering in big data classification," *Inf. Sci.*, vol. 479, pp. 135–152 (2019), doi: 10.1016/j.ins.2018.12.002.

[18] JV. Durga Prasad Jasti, Guttikonda Kranthi Kumar, M. Sandeep Kumar, V. Maheshwari, Prabhu Jayagopal, Bhaskar Pant, Alagar Karthick, and M. Muhibbullah, "Relevant-Based Feature Ranking (RBFR) Method for Text Classification Based on Machine Learning Algorithm," *J. Nanomater.,* vol. 2022, Art. no. 9238968 (2022), doi: 10.1155/2022/9238968.

[19] Nayana Pampapura Madali, Manar Alsaid, and Suliman Hawamdeh, "The impact of social noise on social media and the original intended message: BLM as a case study," *J. Inf. Sci.,* vol. 50, no. 1, pp. 1–17 (2024), doi: 10.1177/01655515221077347.

[20] Yiwei Feng, M. Asif Naeem, Farhaan Mirza, and Ali Tahir, "Reply using past replies — a deep learning-based e-mail client," *Electronics*, vol. 9, no. 9, Art. no. 1353 (2020), doi: 10.3390/electronics9091353.

[21] Belal Abdullah Hezam Murshed, Suresha Mallappa, Jemal Abawajy, Mufeed Ahmed Naji Saif, Hasib Daowd Esmail Al-Ariki, and Hudhaifa Mohammed Abdulwahab, "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis," *Artif. Intell. Rev.,* vol. 56, pp. 5133–5260 (2023), doi: 10.1007/s10462-022-10254-w.

[22] Joel Brynielsson, Fredrik Johansson, Carl Jonsson, and Anders Westling, "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises," *Security Informatics*, vol. 3, Art. no. 7 (2014), doi: 10.1186/s13388-014-0007-3.

[23] M. F. Uddin, "Noise removal and structured data detection," in Proc. ASONAM, pp. 1156–1163 (2016), doi: 10.1109/ASONAM.2016.7752413.

[24] Fatma M. Talaat, Eman M. El-Gendy, Mahmoud M. Saafan, and Samah A. Gamel, "Utilizing social media and machine learning for personality and emotion recognition using PERS," *Neural Comput. Appl.*, vol. 35, no. 33, pp. 23927–23941 (2023), doi: 10.1007/s00521-023-08962-7.

[25] Marco Brambilla, Alireza Javadian Sabet, Kalyani Kharmale, and Amin Endah Sulistiawati, "Graph-based conversation analysis in social media," *Big Data Cogn. Comput.,* vol. 6, no. 4, Art. no. 113 (2022), doi: 10.3390/bdcc6040113.

[26] L. Akoglu, H. Tong, and D. Koutra, "Graph-based anomaly detection and description: A survey," *Data Min. Knowl. Discov.,* vol. 29, no. 3, pp. 626–688 (2015), doi: 10.1007/s10618-014-0365-y.

[27] Jiyoon Lee, Younghoon Kim, and Seoung Bum Kim, "Noise-robust graph-based semi-supervised learning with dynamic shaving label propagation," *Appl. Soft Comput.*, vol. 142, Art. no. 110371, ISSN 1568-4946 (2023), doi: 10.1016/j.asoc.2023.110371.

[28] R. Jain and W. Xu, "AI-based wrapper for high dimensional feature selection," *BMC Bioinformatics,* vol. 24, Art. no. 392 (2023), doi: 10.1186/s12859-023-05502-x.

[29] Simone Pau, Alessandra Perniciano, Barbara Pes, and Dario Rubattu, "An Evaluation of Feature Selection Robustness on Class Noisy Data," Information, vol. 14, no. 8, Art. no. 438 (2023), doi: 10.3390/info14080438.

[30] J. M. Johnson and T. M. Khoshgoftaar, "Classifying big data with label noise: A survey," *ACM J. Data Inf. Qual.,* vol. 14, no. 4, Art. no. 23 (2022), doi: 10.1145/3492546.

[31] Nadhia Salsabila Azzahra, Danang Triantoro Murdiansyah, and Kemas M. Lhaksmana, "Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection," *Int. J. Inf. Commun. Technol.*, vol. 7, no. 1, pp. 64–76 (2021), doi: 10.21108/ijoict.v7i1.552.

[32] T. Zimmerman, "Social noise and information behavior," *J. Doc.,* vol. 78, no. 2, pp. 345–362 (2022), doi: 10.1108/JD-08-2021-0165.

[33] Zahra Nematzadeh, Roliana Ibrahim, and Ali Selamat, "A hybrid model for class noise detection using kmeans and classification filtering algorithms," *SN Appl. Sci.,* vol. 2, Art. no. 1207 (2020), doi: 10.1007/s42452-020-3129-x.

[34] Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Ebrahim Mahdipour, and Seyed Mahdi Jameii, "A systematic review of techniques, open issues, and future directions," *Telemat. Inform.,* vol. 57, Art. no. 101517 (2021), doi: 10.1016/j.tele.2020.101517.

[35] M. Li and C. Zhu, "Noisy label processing for classification: A survey," arXiv preprint arXiv:2404.04159 (2024). [Online]. Available: https://arxiv.org/abs/2404.04159

[36] Kazi Toufique Elahi, Tasnuva Binte Rahman, Shakil Shahriar, Samir Sarker, Md. Tanvir Rouf Shawon, and G. M. Shahariar, "A Comparative Analysis of Noise Reduction Methods in Sentiment Analysis on Noisy Bangla Texts," arXiv preprint arXiv: 2401.14360 (2024). [Online]. Available: https://arxiv.org/abs/2401.14360

[37] A. Saseendran, L. Setia, V. Chhabria, D. Chakraborty and A. Barman Roy, "Impact of Noise in Dataset on Machine Learning Algorithms,"

preprint (2019). [Online]. Available: https://doi.org/10.13140/RG.2.2.25669.91369

[38] L. P. F. Garcia, A. C. Lorena and A. C. P. L. F. Carvalho, "A Study on Class Noise Detection and Elimination," 2012 Brazilian Symposium on Neural Networks, Curitiba, Brazil, pp. 13-18 (2012), doi: 10.1109/SBRN.2012.49.

[39] P. S. Rao, R. N. Sushma, and Anurag, "Study and Analysis of Noise Effect on Big Data Analytics," *Int. J. Manag., Technol. Eng.*, vol. 8, no. 12, pp. 5841–5850 (2018).

[40] R. Alabduljabbar, H. Almazrou, and A. Aldawod, "Context-Aware News Recommendation System: Incorporating Contextual Information and Collaborative Filtering Techniques," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1 (2023). [Online]. Available: https://doi.org/10.1007/s44196-023-00315-5

[41] R. Kaur and R. Maini, "Performance Evaluation and Comparative Analysis of Different Filters for Noise Reduction," *Int. J. Image Graph. Signal Process.*, vol. 8, no. 7, pp. 9–21 (2016). Available :10.5815/ijigsp.2016.07.02

[42] M. Goutay, F. A. Aoudia and J. Hoydis, "Deep Reinforcement Learning Autoencoder with Noisy Feedback," 2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT), Avignon, France, pp. 1-6 (2019), doi: 10.23919/WiOPT47501.2019.9144089.

[43] Rajmadhan Ekambaram, Sergiy Fefilatyev, Matthew Shreve, Kurt Kramer, Lawrence O. Hall, Dmitry B. Goldgof, and Rangachar Kasturi, 'Active cleaning of label noise', Pattern Recognition, vol. 51, pp. 463–480 (2016). Available: https://doi.org/10.1016/j.patcog.2015.09.020

Received July, 2025