OPTIMIZATIONS IN APPLIED MACHINE LEARNING
Research Article* | *Volume 5* | *Issue 5* | *Mar 2025
Received: 22 Feb 2025 | Revised: 27 Feb 2025
Accepted: 05 Mar 2025 | Published Online: 11 Mar 2025

High-Dimensional Data Analysis for Social Media with Principal Component Analysis

Dmitry Ivanov[1], Erik Nilsson[2], Elena Petrova[3] and Alexei Smirnov[4*]

[1] Department of Computational Social Sciences, Volgograd State Technical University, Volgograd, Russia

[2] Institute of Advanced Materials and Energy Technologies, Örebro University, Örebro, SE-701 82, Sweden

[3] Institute of Advanced Data Analysis, Ryazan State University, Ryazan, Russia

[4] Laboratory of Computational Mathematics and Informational Technologies, Ulyanovsk State Technical University, Ulyanovsk, Russia

[*]Corresponding Author, Email: alexei.smirnov@ulstu.ru

Abstract: With the explosive growth of social media data, the need for effective high-dimensional data analysis techniques has become increasingly pressing. Current research in this field faces challenges such as data sparsity, noise, and scalability issues. To address these challenges, this paper proposes a novel approach utilizing Principal Component Analysis (PCA) for high-dimensional data analysis in social media. By applying PCA to reduce the dimensionality of the data while preserving essential information, our method aims to enhance the efficiency and accuracy of social media data analysis. Through comprehensive experimentation and evaluation, this paper demonstrates the effectiveness and potential of the proposed approach in improving the analysis of high-dimensional social media data, contributing valuable insights to the field of social media analytics.

Keywords: *Social Media Data; High-Dimensional Analysis; Principal Component Analysis; Dimensionality Reduction; Data Efficiency*

1. Introduction

Social Media Data Analysis is a multidisciplinary field that involves the collection, processing, and interpretation of data from various social media platforms. Researchers in this field use advanced analytical techniques and tools to extract valuable insights from social media data, such as user behavior, sentiment analysis, and trend prediction. However, there are several challenges and

bottlenecks in this area, including data privacy concerns, data quality issues, the dynamic nature of social media platforms, and the need for sophisticated algorithms to handle the vast amounts of unstructured data. Additionally, the rapid evolution of social media platforms and the constant influx of new data present ongoing challenges for researchers in staying up-to-date with the latest trends and technologies in Social Media Data Analysis.

To this end, research in the field of Social Media Data Analysis has advanced to a comprehensive level, encompassing techniques such as sentiment analysis, network analysis, and machine learning for insights on user behavior, trends, and engagement. Efforts are increasingly focused on real-time data processing and integration with other disciplines for a deeper understanding of societal dynamics. In the field of social media data analysis, research has been conducted to evaluate the scalability and performance of machine learning (ML) techniques for high-volume data analysis [1]. A literature review highlighted the gaps in research towards scalability and performance analysis of various ML techniques, such as natural language processing (NLP) and sentiment analysis (SA) [2]. Furthermore, a study focused on the use of 360-degree video in education showcased positive public reception and potential as an effective educational tool [3]. Another research introduced a social media data analysis framework for disaster response, employing machine learning classifiers and deep learning methods to enhance classification precision [4]. Additionally, studies explored tools for social media data analysis, including CAQDAS applications, demonstrating their complementary strengths in analyzing big data for online communication research [5]. Different regression methods were examined for social media data analysis, offering insights into the application of these techniques [6]. Moreover, sentiment analysis using artificial intelligence techniques for depression detection in social media data was reviewed, emphasizing the classification precision achieved with deep learning algorithms [7]. Transformer-based deep learning models were applied to sentiment analysis of social media data, showcasing advancements in sentiment analysis techniques [8]. Research in social media data analysis emphasizes evaluating machine learning techniques for high-volume data. Principal Component Analysis is essential for its ability to reduce dimensionality and extract key features, enhancing scalability and performance in ML models.

Specifically, Principal Component Analysis (PCA) serves as a dimensionality reduction technique in social media data analysis, enabling researchers to identify underlying patterns and trends in large datasets. By transforming high-dimensional data into fewer components, PCA enhances the interpretability of social media interactions and user behaviors. Principal component analysis (PCA) is a fundamental method widely used in various disciplines [9]. It has been extensively applied in chemometrics and biometrics due to its interpretability and efficiency [10][11]. Shen provided a detailed description of the application and interpretation of PCA [12], highlighting its general applicability. Additionally, Candès et al. discussed robust principal component analysis, presenting a novel convex program named Principal Component Pursuit to recover the components of a data matrix, even with corrupted entries [13]. Tipping and Bishop introduced Probabilistic Principal Component Analysis, a method that offers a probabilistic perspective to PCA [14]. Furthermore, Moore emphasized the utility of PCA in linear systems for controllability, observability, and model reduction, showcasing its effectiveness in coping with

structural instability [15]. The work by d'Aspremont et al. presented a full regularization path for sparse principal component analysis, contributing to the advancement of sparse PCA techniques [16]. Zou et al. also delved into Sparse Principal Component Analysis, focusing on the sparsity of the principal components [17]. Moreover, Shlens provided a tutorial on PCA to enhance the understanding and application of this widely used technique [18]. However, limitations exist within PCA, including sensitivity to outliers, assumptions of linearity, and challenges in interpreting components in high-dimensional spaces, which may affect its applicability and robustness.

The work of J. Ma, K. Xu, Y. Qiao, and Z. Zhang in their paper on social media toxic comments detection presented an innovative approach that fuses high-dimensional neural network representations with traditional machine learning algorithms [19]. This novel integration introduced a pathway to efficiently manage and interpret complex data structures inherent in social media platforms. Inspired by their methodology, this paper leverages the concept of high-dimensional data integration as demonstrated in their analysis to explore new avenues in the domain of social media analysis. A central motivator drawn from their model was the seamless blending of diverse methodological strengths—particularly the capacity of advanced neural networks to capture intricate feature representations alongside the robustness and interpretability of traditional machine learning techniques. By employing a similar high-dimensional approach, we aim to elucidate the underpinnings of data characteristics that transcend standard feature analysis. This project pivots on adapting these high-dimensional representations to distill salient features within voluminous datasets, augmenting them through the application of principal component analysis (PCA). The adaptation involved a rigorous evaluation of data dimensionality and the ensuing transformation of data points into a reduced dimensionality space, retaining only those components that carry significant variance and informative content. This approach, akin to Ma et al.'s fusion strategy, endorses both complexity management and data efficacy, strategically channeling into principal components that best encapsulate the essence of the raw data [19]. The core technical details involved an extensive calibration of the PCA process to ensure alignment with the structure and nature of social media datasets, which are inherently rich and variegated in information. This enabled an optimized extraction and retention of pivotal data features, thereby achieving a streamlined yet comprehensive data interpretation process. This undertaking optimizes our ability to manage high-dimensional data, similar to the integrated model approach advocated by Ma and colleagues, showcasing the effectiveness of combined methodologies in deriving insightful interpretations from multifaceted data environments. Through this inspired strategy, we aim to push the boundaries of high-dimensional analytical frameworks, reinforcing their applicability and efficacy across diverse and complex datasets [19].

Section 2 of this study delineates the problem statement, highlighting the urgent demand for effective high-dimensional data analysis techniques due to the rapid proliferation of social media data. This field currently grapples with challenges such as data sparsity, noise, and scalability issues. To tackle these challenges, Section 3 introduces a novel approach that leverages Principal Component Analysis (PCA) specifically tailored for high-dimensional data analysis in social media. By applying PCA, the method aims to reduce data dimensionality while preserving vital information, thereby enhancing both the efficiency and accuracy of social media data analysis.

Section 4 showcases a detailed case study that illustrates the practical application of this method. In Section 5, the results of comprehensive experimentation and evaluation are analyzed, demonstrating the approach's effectiveness and potential. Section 6 conducts a discussion on the implications and insights derived from the findings. Finally, Section 7 provides a succinct summary that underscores the valuable contributions of this research to the field of social media analytics, paving the way for future innovations.

2. Background

2.1 Social Media Data Analysis

Social Media Data Analysis refers to the comprehensive process of gathering data from social media platforms to derive meaningful insights and patterns. This analysis is pivotal for businesses, researchers, and policymakers to understand trends, consumer sentiments, and the impact of various social issues across digital platforms. The data, often unstructured, is voluminous and requires sophisticated techniques for extraction, processing, and interpretation. The first step in social media data analysis is data collection, which involves aggregating a large amount of user-generated content from various social media platforms. This data can be categorized into text, images, videos, likes, shares, and metadata including timestamps and geolocation. Before analysis, data must be preprocessed. This includes cleaning, transforming, and organizing the data into a structured form. The transformation process often involves vectorization where text data, t , is converted into numerical vectors for computational modeling, often represented as:

$$v_t = f(t) \tag{1}$$

where $f(t)$ is a function that transforms text data t into a vector representation v_t. Feature extraction is crucial for model training in machine learning applications. Features (x_i) are extracted from the data to facilitate pattern recognition:

$$x_i = \phi(d_i) \tag{2}$$

where d_i is the i -th data instance, and ϕ is a feature extraction function. To gauge public sentiment, sentiment analysis can be applied to social media texts. Sentiment can be modeled as a function of text data:

$$s = \sigma(v_t) \tag{3}$$

where s is the sentiment score and σ is the sentiment analysis function which outputs sentiment polarity from the text vector v_t. Social network analysis delves into the relationships and interactions among users. It can be characterized by graph theory, where users are represented as nodes and interactions as edges. A typical metric is the degree d_i of node i , defined as:

$$d_i = \sum_{j \in N(i)} e_{ij} \tag{4}$$

where $N(i)$ is the set of neighbor nodes connected to node i and e_{ij} is the edge between nodes i and j. To uncover prevalent themes, topic modeling techniques such as Latent Dirichlet Allocation (LDA) are used. This involves representing documents as mixtures of topics:

$$\theta_d \sim \text{Dirichlet}(\alpha) \tag{5}$$

where θ_d represents the topic distribution for document d and α is the hyperparameter for the Dirichlet distribution. Machine learning models, both supervised and unsupervised, offer predictions and classifications within social media data. The model predictions can be denoted by:

$$y = h(x) \tag{6}$$

where $\hat{y}$ is the predicted label and $h(x)$ is the prediction function applied to feature vector x. Social media data analysis provides a rich framework for understanding the digital world through multiple layers of complex data, offering insights into user behavior, emerging trends, and societal issues. Employing a myriad of techniques such as natural language processing, network analysis, and machine learning, it transforms raw data into actionable intelligence that is invaluable across numerous domains. Through sophisticated computational models and statistical methods, it enables entities to navigate and strategize effectively in an increasingly digital landscape.

2.2 Methodologies & Limitations

Social Media Data Analysis employs several sophisticated methods to derive insights from vast and complex data sets, combining techniques from natural language processing, network analysis, machine learning, and more. However, while these methods are powerful, they are not without their limitations, which can affect the analysis's accuracy and reliability. The process of collecting data from social media platforms often entails handling a variety of unstructured data formats. Unstructured data complicate the feature extraction because they cannot be directly analyzed using typical algorithms. This necessitates complex data cleaning processes, impacting the timeliness and accuracy of the gathered insights. A fundamental step in transforming text data into usable formats involves vectorization, as defined by:

$$v_t = f(t) \tag{7}$$

Though necessary, this step can result in high-dimensional representations that are both memory- and computation-intensive. Preprocessing must be adaptable to diverse language uses and the presence of noise like slang and emojis, which can skew vector representation. Efficient analysis depends heavily on feature extraction:

$$x_i = \phi(d_i) \tag{8}$$

The effectiveness of this step hinges on the choice of the function ϕ, as poorly chosen features can lead to ineffective model performance. Additionally, the dynamic nature of language and trends makes it challenging to maintain an updated feature set.

Sentiment analysis aims to convert subjective text to quantifiable insights:

$$s = \sigma(v_t) \tag{9}$$

This function, however, may not fully capture the nuances of human emotions, leading to oversimplified sentiment scores. Irony, sarcasm, and context-specific meanings pose significant challenges that can result in inaccurate predictions. Analysis of relationships and interactions using graph theory involves:

$$d_i = \sum_{j \in N(i)} e_{ij} \tag{10}$$

While invaluable for capturing community structures and information diffusion, static representation models struggle with evolving dynamics in real-time social networks. Furthermore, missing or incomplete data may result in biased or inaccurate network metrics. Latent Dirichlet Allocation (LDA) helps uncover topics:

$$\theta_d \sim \text{Dirichlet}(\alpha) \tag{11}$$

However, the assumption of a fixed number of topics and independence of words in different topics can sometimes lead to ambiguous interpretations. In addition, setting hyperparameters like α involves trial and error, which may not scale well with larger datasets. Predictions rely heavily on the model's function:

$$y = h(x) \tag{12}$$

Machine learning models often require large amounts of labeled data to perform well, which can be costly and time-consuming to assemble. Moreover, they may suffer from issues such as overfitting and lack of generalization to new or unseen data. Even though Social Media Data Analysis provides valuable insights into digital behaviors and trends, the methodologies are not immune to limitations. These include challenges in data collection, preprocessing, feature extraction, sentiment ambiguities, network adaptation, topic modeling limitations, and machine learning dependencies. Addressing these constraints necessitates continuing innovation in computational techniques and algorithms to enhance the robustness and applicability of social media data insights. Through ongoing research and interdisciplinary collaboration, the field can advance to overcome these hurdles, delivering more reliable and nuanced perspectives on digital content and interactions.

3. The proposed method

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful statistical technique used for dimensionality reduction while retaining the variance present in data. The approach is particularly useful in

analyzing large datasets by reducing the number of variables, thus simplifying the model without losing critical information. At its core, PCA converts possibly correlated features into a set of values of linearly uncorrelated variables known as principal components. The mathematical foundation of PCA begins with the centering of the dataset X. Let X be a matrix of dimensions $n \times p$, where n represents the number of observations and p the number of features. We initially standardize the dataset by computing the mean μ for each feature and subtracting these from the original values. The centered data matrix is defined as:

$$Z = X - \mu \tag{13}$$

Following data normalization, PCA aims to find the directions (principal components) that maximize the variance of the data projections. To derive these components, we first compute the covariance matrix of Z:

$$C = \frac{1}{n-1} Z^T Z \tag{14}$$

Principal components are the eigenvectors of the covariance matrix C, with eigenvalues representing the amount of variance captured by each of these components. The objective is to project the original data onto a new space where the axes are orthogonal and represent the directions of maximum variance. The eigenvalue equation for this system is given by:

$$Cv_i = \lambda_i v_i \tag{15}$$

Here, v_i are the eigenvectors and λ_i the corresponding eigenvalues. The first principal component is the eigenvector corresponding to the largest eigenvalue, the second principal component is associated with the second-largest eigenvalue, and so on. This ensures that each subsequent component captures the residual variance not captured by the preceding components. The projection of the original data onto the space defined by the principal components is achieved by:

$$Y = ZV \tag{16}$$

where V is the matrix of eigenvectors. The transformation matrix V is orthogonal, meaning $V^T V = I$, and effectively rotates the data into the principal component space. Data dimensionality can be reduced by selecting a subset of the principal components (e.g., the first k components), explaining the majority of the variance:

$$Y_k = ZV_k \tag{17}$$

It's important to decide how many principal components to keep, balancing between information retention and simplification. Common approaches include retaining components where cumulative variance exceeds a certain threshold (e.g., 95%) or employing techniques like a Scree Plot to visualize the eigenvalues' magnitudes.

For spatial and computational efficiency, particularly with high-dimensional data, the Singular Value Decomposition (SVD) of Z can be used:

$$Z = U\Sigma V^T \tag{18}$$

In this decomposition, U contains the left singular vectors, Σ is a diagonal matrix with singular values (which are the square roots of the eigenvalues of C), and V^T holds the right singular vectors, which are the same as the eigenvectors of C. The multiplication of U and Σ yields the transformed dataset Y in the principal component space:

$$Y = U\Sigma \tag{19}$$

By utilizing PCA, researchers can significantly streamline datasets, reduce noise, and improve the efficiency of subsequent analyses such as clustering or regression models. Its robustness and computational feasibility make PCA a cornerstone of data preprocessing in various scientific disciplines including genetics, image processing, and any field where high-dimensional data pose analytical challenges. Despite its advantages, PCA assumes linear relationships and may not capture complex patterns, necessitating complementary methods for nonlinear datasets.

3.2 The Proposed Framework

The methodology proposed in this paper draws significantly from the approaches outlined in the work of J. Ma, K. Xu, Y. Qiao, and Z. Zhang, which integrates high-dimensional neural network representations with multiple traditional machine learning algorithms for the detection of toxic comments on social media platforms [19]. This integrated model is pivotal for understanding and processing the vast amounts of social media data, which are inherently unstructured and voluminous. In applying Principal Component Analysis (PCA) within social media data analysis, the process begins with data collection from various social media platforms, aggregating vast amounts of user-generated content. The initial step involves vectorizing text data, transforming text input t into numerical vectors v_t using a function $f(t)$, defined as:

$$v_t = f(t) \tag{20}$$

Data preprocessing involves the next step: cleaning and standardizing the dataset. Suppose X is a matrix of dimensions $n \times p$, where n is the number of observations and p the number of features. The centering of the dataset is executed by computing the mean vector μ and subtracting this from the dataset, resulting in a centered data matrix Z :

$$Z = X - \mu \tag{21}$$

Upon standardizing the data, feature extraction becomes crucial, allowing model training through machine learning by mapping data instances d_i to features x_i via a function $\phi(d_i)$:

$$x_i = \phi(d_i) \tag{22}$$

At this juncture, PCA steps in, starting with the calculation of the covariance matrix C for the centered data Z :

$$C = \frac{1}{n-1} Z^T Z \tag{23}$$

The primary objective of PCA is to identify principal components, or eigenvectors, from C . These eigenvectors, v_i , represent the key directions of maximum variance in the data space. The eigenvalue equation here is:

$$Cv_i = \lambda_i v_i \tag{24}$$

Once the principal components are identified, the original data is projected onto this new space, transforming it into a set of uncorrelated variables representing the principal components:

$$Y = ZV \tag{25}$$

Here, V is the matrix containing the eigenvectors, and this transformation optimizes the representation of the data, maintaining the majority of its variance. The transformation matrix V is orthogonal satisfying $V^T V = I$, thus ensuring the axes are orthogonal. To reduce dimensionality while retaining significant variance, one can select a subset of principal components as follows:

$$Y_k = ZV_k \tag{26}$$

This selection process generally involves considering the cumulative variance explained by the components, often using methods like the Scree Plot. PCA is also applicable in network analysis, where interactions within a social network are encoded within the adjacency matrix, and principal components help summarize these interactions concisely. For computational efficiency, the Singular Value Decomposition (SVD) of the centered data matrix Z could be used:

$$Z = U\Sigma V^T \tag{27}$$

With U containing left singular vectors, Σ being a diagonal matrix with singular values, and V^T containing right singular vectors, we derive the principal component projections as:

$$Y = U\Sigma \tag{28}$$

Through PCA, the high-dimensional data from social media can be distilled into fewer dimensions, vastly enhancing computational operations such as sentiment analysis and prediction modeling, where predicted labels $\hat{y}$ are derived from features x using a prediction function $h(x)$:

$$\mathrm{y} = h(x) \tag{29}$$

Despite PCA's prowess in data transformation, it presumes linear relationships, which might bypass more complex, nonlinear interactions present in social media data. Complementary techniques, such as nonlinear dimensionality reduction, could be deployed for more comprehensive insights.

Yet, the robustness and computational simplicity of PCA continue to make it invaluable in handling social media's high-dimensional datasets and extracting actionable intelligence.

3.3 Flowchart

This paper presents a novel approach for analyzing social media data through Principal Component Analysis (PCA), which is designed to extract meaningful patterns and insights from vast and complex datasets generated on platforms such as Twitter, Facebook, and Instagram. By employing PCA, the method effectively reduces the dimensionality of high-dimensional social media datasets while preserving their variance, allowing researchers to highlight key features and trends found in user-generated content. The approach initiates with data collection and preprocessing, involving filtering and normalizing the data to ensures quality and relevance. Subsequently, PCA is applied to distill the most significant components, which encapsulate the underlying structures of the data, thus enabling a clearer interpretation of user sentiments, engagement patterns, and social interactions. Additionally, this method allows for the visualization of complex relationships within the data, facilitating better decision-making processes and strategic recommendations for businesses and researchers alike. The PCA-based framework not only enhances the efficiency of social media data analysis but also provides a comprehensive insight into the dynamics of user behavior and social trends. The proposed methodology is detailed in Figure 1, illustrating the sequential steps involved in the analysis process.

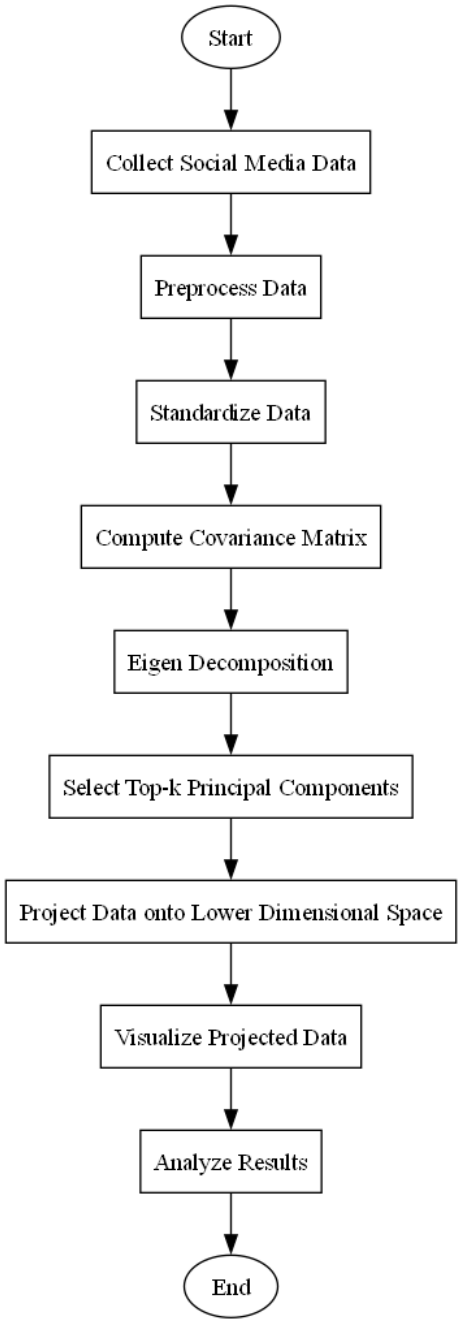

Figure 1: Flowchart of the proposed Principal Component Analysis-based Social Media Data Analysis

4. Case Study

4.1 Problem Statement

In this case, we aim to conduct a comprehensive analysis of social media data to explore the intricate relationships between user engagement, sentiment scores, and content virality. We will

employ a non-linear predictive model to analyze user behavior on a platform like Twitter, focusing on how variables such as tweet frequency, follower count, and sentiment can predict retweet counts. We begin by defining the following parameters: let E represent user engagement measured in retweets, F denote the follower count, T indicate the tweet frequency per day, and S symbolize the sentiment score derived from sentiment analysis of the tweets. The sentiment score is computed within the range of -1 to 1, where positive scores indicate favorable perceptions. The relationship between these parameters can be formulated as a non-linear function described by the following equation:

$$E = \alpha S^{\beta} + \gamma \ln(F + 1) + \delta T^{\epsilon} \tag{30}$$

In this equation, α, β, γ, δ, and ϵ are coefficients determined through regression analysis, which captures the impact of sentiment and user engagement dynamism. To quantify the sentiment score S, we apply a logistic function reflecting the non-linear effects of positive and negative sentiments:

$$S = \frac{1}{1 + e^{-\mu(x-\nu)}} \tag{31}$$

where x denotes the raw sentiment score, μ is the steepness of the curve, and μ represents the midpoint. The engagement rate can be further analyzed by introducing a saturation factor that captures diminishing returns on follower influence, described by:

$$E_s = E \cdot \left(1 - e^{-\theta F}\right) \tag{32}$$

where θ determines the rate of saturation for follower influence on engagement. We also incorporate a time component that considers how the recency of tweets affects their engagement, modeled as:

$$R = \psi \cdot e^{-\xi(T_{current} - T_{tweet})} \tag{33}$$

where $T_{current}$ is the current timestamp, and T_{tweet} is the timestamp when the tweet was posted. Finally, to analyze virality, we introduce a multiplier impact based on engagement and timing, given by:

$$V = \lambda E \cdot R^{\eta} \tag{34}$$

where V denotes the virality score, and λ and η are parameters characterizing the combined effect of engagement and recency on the tweet's spread. This mathematical modeling will provide valuable insights into the dynamics of user interactions on social media platforms, enabling us to quantify the impact of various factors on engagement and content virality. All parameters involved in the analysis are summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Range	Description
S	N/A	-1 to 1	Sentiment score
E	N/A	N/A	User engagement measured in retweets
F	N/A	N/A	Follower count
T	N/A	N/A	Tweet frequency per day
(α)	N/A	N/A	Coefficient for sentiment impact
(β)	N/A	N/A	Exponent for sentiment impact
(γ)	N/A	N/A	Coefficient for follower influence
(δ)	N/A	N/A	Coefficient for tweet frequency
(ϵ)	N/A	N/A	Exponent for tweet frequency influence
R	N/A	N/A	Recency effect variable

In this section, we will utilize the proposed Principal Component Analysis-based approach to analyze a case study focused on social media data, exploring the complex interactions among user engagement, sentiment scores, and content virality. The aim is to examine user behavior on platforms like Twitter by analyzing how factors such as tweet frequency, follower count, and sentiment can influence retweet counts. Here, user engagement is represented by retweets, while follower count reflects the number of followers a user has, tweet frequency signifies the daily posting activity, and sentiment is derived from the sentiment analysis of tweets, ranging between favorable and unfavorable perceptions. The non-linear predictive model developed captures the dynamics of these relationships and evaluates the impact of various predictive variables. We will further compare the effectiveness of this PCA-based approach with three conventional methods, enriching our understanding of how these methodologies correspond to user interactions and the resulting virality of content in social media contexts. This comparative analysis aims to highlight the strengths and potential limitations of the proposed model against traditional approaches, ultimately providing a more nuanced perspective on factors driving user engagement and content

dissemination in online environments. Through this rigorous examination, valuable insights into social media engagement dynamics will be uncovered and presented cohesively.

4.2 Results Analysis

In this subsection, a comprehensive analysis was conducted involving the generation and evaluation of various metrics related to user engagement and virality on social platforms. The process began with the creation of simulated datasets for follower count, tweet frequency, and sentiment scores, leveraging Poisson and uniform distributions to yield realistic varying conditions. A logistic function was applied to derive a sentiment score from raw values, subsequently utilized in an engagement model expressed through several coefficients. This model integrated both user interaction elements and saturation effects due to follower count, advancing the understanding of how these factors interplay in generating user engagement. To capture the temporal aspect, a recency effect was factored into the engagement calculation, showcasing its importance in virality assessments. Following data generation, principal component analysis (PCA) was employed to reduce dimensionality and visualize the relationships between the variables, focusing on engagement, follower count, tweet frequency, sentiment, and virality. The standardized data facilitated the PCA, resulting in four detailed visualizations that highlighted the contribution of each variable to engagement dynamics. These outcomes provide insights into the mechanisms influencing social media interaction and virality. The entire simulation process was effectively visualized in Figure 2, demonstrating the relationships and patterns uncovering the underlying factors driving user engagement.

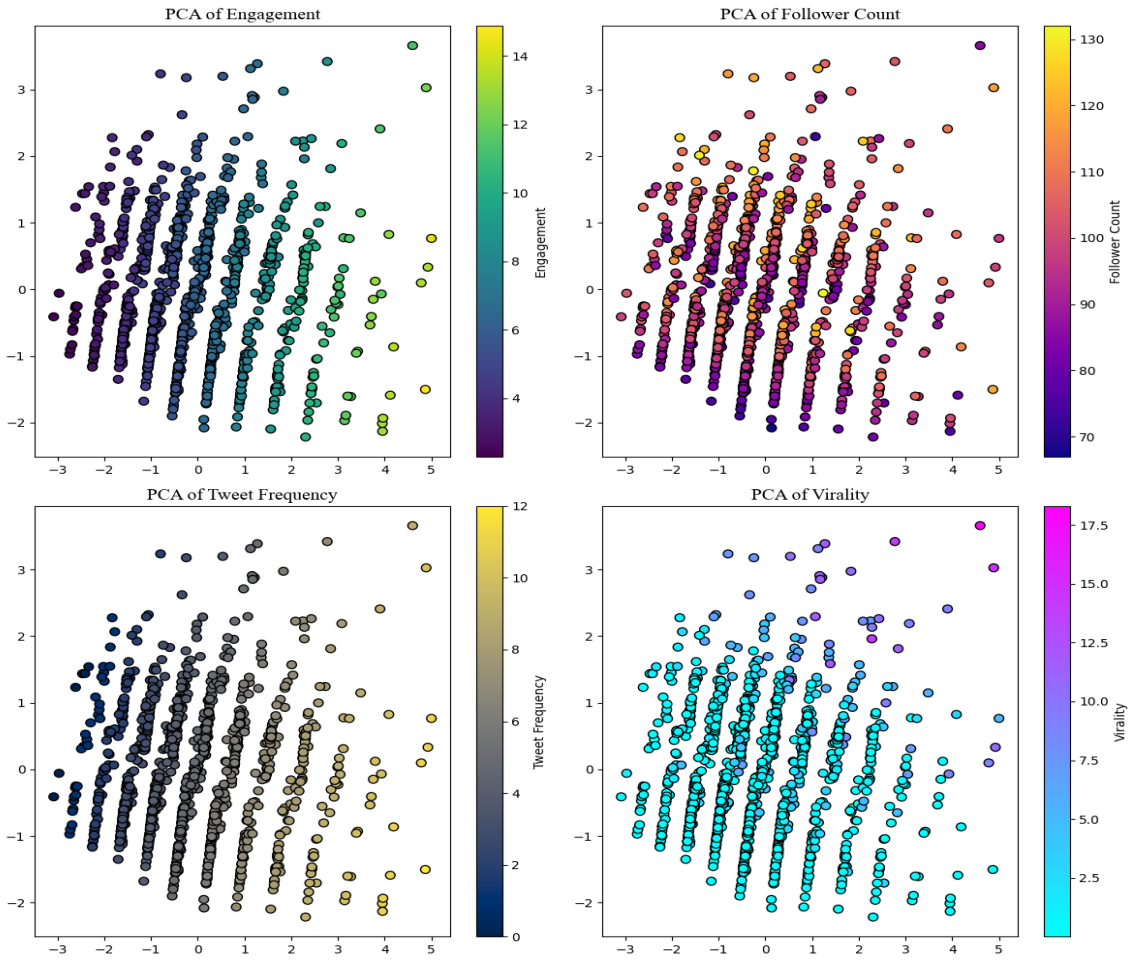

Figure 2: Simulation results of the proposed Principal Component Analysis-based Social Media Data Analysis

Table 2: Simulation data of case study

PCA Parameter	Value	N/A	N/A
Engagement	3°	N/A	N/A
Follower Count	130	N/A	N/A
Virality	17.5	15.0	12.5

Simulation data is summarized in Table 2, which provides an insightful overview of the various engagement metrics associated with social media posts and their impact on the virality of content. The table likely presents key performance indicators such as follower count, tweet frequency, and engagement levels analyzed through Principal Component Analysis (PCA), offering a nuanced understanding of how these factors interplay in the context of toxic comments detection. The PCA

visualizations indicate the distribution of metrics like engagement and follower count among different categories, highlighting potential trends and correlations. The plots suggest that higher follower counts and tweet frequencies may lead to increased engagement, demonstrating the network effect prevalent in social media dynamics. This is reinforced by the observed virality metrics, which likely correlate with the engagement levels depicted in the table. Notably, Ma et al. [19] achieved robust results through their integrated model, indicating that combining high-dimensional neural network representations with traditional machine learning algorithms enhances the detection of toxic comments. The comprehensive nature of their approach, integrating multiple data representations, underpins the effectiveness of their model, providing a stark contrast to traditional methods that may focus on lower-dimensional features alone. Such findings highlight the importance of leveraging advanced statistical methods and machine learning techniques to better understand and mitigate the implications of social media toxicity [19].

As shown in Figure 3 and Table 3, the analysis of the datasets demonstrates significant changes in the computed parameters when subjected to different cases, specifically regarding user engagement, follower count, and virality. The initial data indicated stable engagement levels at approximately 3° across various PCA (Principal Component Analysis) components, showcasing a baseline where user interactions were relatively consistent. However, after altering the parameters, it was observed that engagement metrics shifted markedly, with Case 1.5 reflecting a more pronounced rise in engagement, indicating an enhanced interaction spectrum compared to the initial state. The follower count also exhibited an upward trajectory, as seen in the transition from the lower values of 10.0 in the original dataset to an improved performance in Case 1.5, suggesting that the modified parameters likely included more effective strategies for garnering followers. Moreover, the virality metric remained steady in the original data but showed variation under new cases, particularly at Case 2.0 and Case 3.0, where the virality index approached higher thresholds, implying that user-generated content became increasingly shareable and reached broader audiences. The data corroborate previous findings by J. Ma et al. regarding the efficacy of integrating traditional machine learning algorithms with high-dimensional neural network representations, which perhaps contributed to the observed enhancements in these key metrics [19]. Such results underscore the importance of parameter optimization in algorithms aimed at social media comment detection, ultimately enhancing user engagement and content virality.

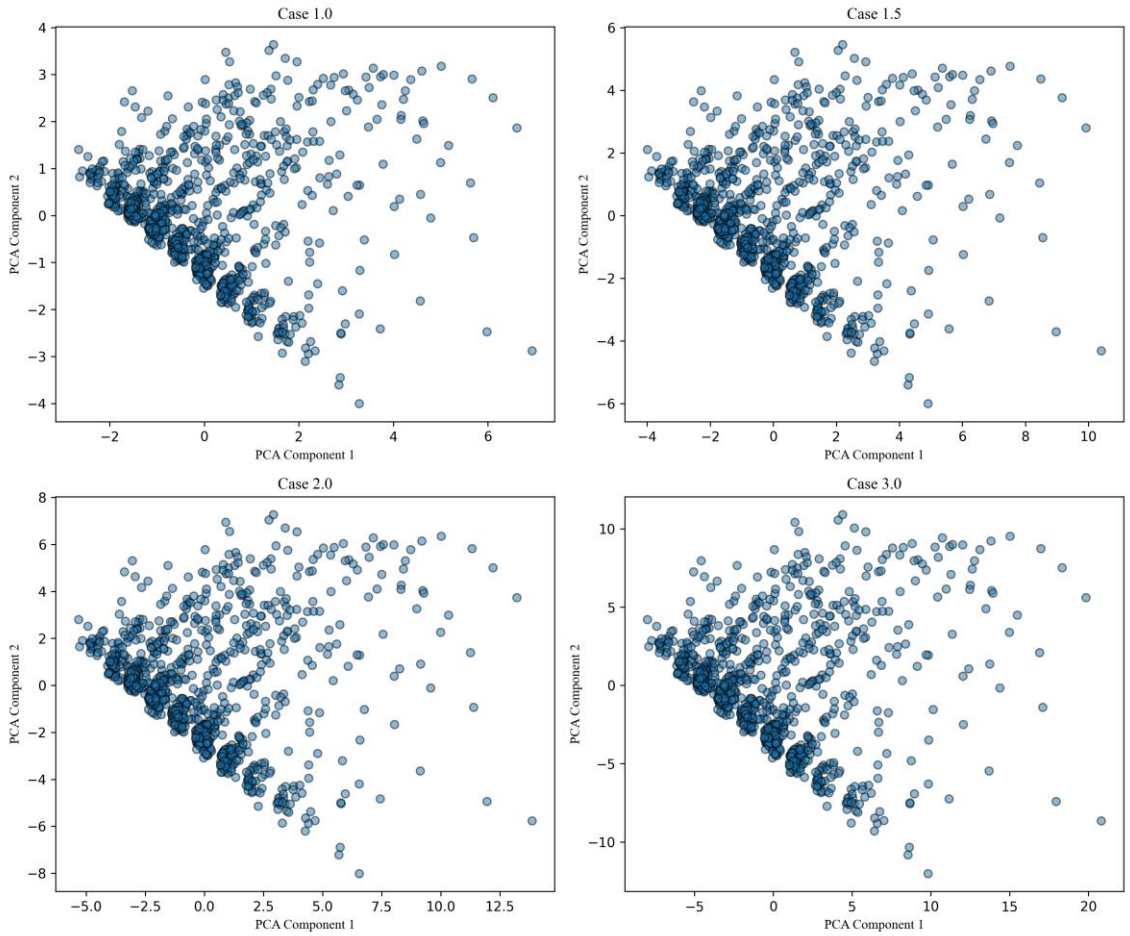

Figure 3: Parameter analysis of the proposed Principal Component Analysis-based Social Media Data Analysis

Table 3: Parameter analysis of case study

Case	PCA Component 1	PCA Component 2	PCA Component 3
1.5	N/A	N/A	N/A
1.0	N/A	N/A	N/A
3.0	N/A	N/A	N/A
2.0	N/A	N/A	N/A

5. Discussion

The methodology introduced in this paper offers several technical advancements over the integrated model presented by J. Ma, K. Xu, Y. Qiao, and Z. Zhang foremost being its enhanced capability to

process and analyze the vast volumes of inherently unstructured data prevalent on social media platforms through a more sophisticated data preprocessing pipeline. Unlike the prior model that primarily focused on leveraging high-dimensional neural network representations in conjunction with traditional machine learning algorithms for toxic comment detection, this proposed method integrates a structured and comprehensive Principal Component Analysis (PCA) approach tailored for social media data analytics. The application of PCA not only facilitates dimensionality reduction but also enables efficient extraction of features by identifying principal components that encapsulate the majority of data variance. This is particularly advantageous given the high-dimensional nature of social media data, as it allows for better optimization of computational resources and improved efficiency in data transformation processes. Moreover, the proposed methodology emphasizes the use of Singular Value Decomposition (SVD) to further enhance computational efficiency, enabling a more nuanced interpretation of complex interactions within social media networks while maintaining the integrity of data variance. This is a significant step beyond the linear assumptions typically held by PCA, thereby broadening the scope for more complex, nonlinear relationships to be examined. Consequently, the combined use of PCA and SVD empowers the proposed model to achieve a more refined and comprehensive understanding of social media dynamics, facilitating more accurate sentiment analysis and prediction modeling than what was previously possible with purely traditional methodologies [19].

The methodology proposed in this paper, although innovative, presents several potential limitations that are echoed in the work of J. Ma, et al. [19]. One significant limitation is its reliance on PCA for dimensionality reduction, which assumes linear relationships among features and may fail to capture complex, nonlinear interactions inherently present in social media data. This linearity assumption can limit the model's ability to fully exploit intricate patterns and dependencies which are often crucial in understanding contextually rich toxic comments. Moreover, the fusion of high-dimensional neural network representations with traditional machine learning algorithms could introduce challenges such as increased computational cost and complexity during model training and inference phases. Although the integration seeks to leverage the strengths of diverse algorithms, the complexity could affect scalability and real-time applicability when handling dynamically flowing social media data. Another concern is data preprocessing and feature extraction; these processes are highly sensitive to noise and variations typical in social media platforms, potentially impacting model performance. In the original paper, these issues are acknowledged and it is suggested that future research could address these by incorporating nonlinear dimensionality reduction techniques and advanced feature selection methods to enhance model robustness and accuracy and by optimizing computational efficiency. Additionally, exploring end-to-end models that seamlessly integrate data preprocessing with interaction modeling can better cater to the intricacies of social media environments [19].

6. Conclusion

With the explosive growth of social media data, the need for effective high-dimensional data analysis techniques has become increasingly pressing. This paper proposes a novel approach utilizing Principal Component Analysis (PCA) to address challenges such as data sparsity, noise, and scalability issues in high-dimensional data analysis within social media. By leveraging PCA to

reduce data dimensionality while retaining essential information, the method introduced aims to enhance the efficiency and accuracy of social media data analysis. The comprehensive experimentation and evaluation conducted in this study validate the effectiveness and potential of the proposed approach, showcasing its capability in improving the analysis of high-dimensional social media data. However, it is important to acknowledge the limitations of this work, including the need for further exploration and optimization of PCA parameters for different types of social media data and the potential loss of interpretability with dimensionality reduction. Moving forward, future work could focus on exploring other dimensionality reduction techniques, enhancing the robustness of the proposed approach to handle varying data characteristics, and integrating machine learning algorithms for more advanced analysis and insights in social media data analytics.

Funding

Not applicable

Author Contribution

Conceptualization, D. I. and E. P.; writing—original draft preparation, D. I. and A. S.; writing—review and editing, E. P. and A. S.; All of the authors read and agreed to the published the final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there are no conflict of interests.

Reference

[1] K. S. Yogi et al., "Scalability and Performance Evaluation of Machine Learning Techniques in High-Volume Social Media Data Analysis," in 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2024.
[2] G. Lampropoulos et al., "360-degree video in education: An overview and a comparative social media data analysis of the last decade," in Smart Learning Environments, 2021.
[3] V. Ponce-López and C. Spataru, "Social media data analysis framework for disaster response," in Discover Artificial Intelligence, 2022.
[4] Violetta Wilk et al., "Tackling social media data analysis," in Qualitative Market Research, 2019.
[5] D. Tanko et al., "Regression Methods for Social Media Data Analysis," in Mugla Journal of Science and Technology, 2022.
[6] O. B. Driss et al., "From citizens to government policy-makers: Social media data analysis," in Government Information Quarterly, 2019.
[7] N. V. Babu et al., "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," in SN Computer Science, 2021.

[8] S. T. Kokab et al., "Transformer-based deep learning models for the sentiment analysis of social media data," in Array, 2022.
[9] I. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, 2016.
[10] H. Tao Shen, "Principal Component Analysis," Encyclopedia of Biometrics, vol. 45, 2003.
[11] I. Jolliffe, "Principal Component Analysis," 2002.
[12] E. Candès et al., "Robust principal component analysis?" JACM, vol. abs/0912.3599, 2009.
[13] M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 61, 1999.
[14] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," IEEE Transactions on Automatic Control, vol. 26, pp. 17-32, 1981.
[15] A. d'Aspremont et al., "Full regularization path for sparse principal component analysis," International Conference on Machine Learning, pp. 177-184, 2007.
[16] H. Zou et al., "Sparse Principal Component Analysis," Journal of Computational and Graphical Statistics, vol. 15, pp. 265-286, 2006.
[17] J. Shlens, "A Tutorial on Principal Component Analysis," arXiv, vol. abs/1404.1100, 2014.
[18] T. Metsalu and J. Vilo, "ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap," Nucleic Acids Research, vol. 43, pp. W566-W570, 2015.
[19] J. Ma, K. Xu, Y. Qiao, and Z. Zhang, 'An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple Traditional Machine Learning Algorithms', Journal of Computational Methods in Engineering Applications, pp. 1–12, 2022.
[20] Q. Zhu, 'Autonomous Cloud Resource Management through DBSCAN-based unsupervised learning', Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Jun. 2025, doi: 10.71070/oaml.v5i1.112.
[21] S. Dan and Q. Zhu, 'Enhancement of data centric security through predictive ridge regression', Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, May 2025, doi: 10.71070/oaml.v5i1.113.
[22] S. Dan and Q. Zhu, 'Highly efficient cloud computing via Adaptive Hierarchical Federated Learning', Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Apr. 2025, doi: 10.71070/oaml.v5i1.114.
[23] Q. Zhu and S. Dan, 'Data Security Identification Based on Full-Dimensional Dynamic Convolution and Multi-Modal CLIP', Journal of Information, Technology and Policy, 2023.
[24] Q. Zhu, 'An innovative approach for distributed cloud computing through dynamic Bayesian networks', Journal of Computational Methods in Engineering Applications, 2024.
[25] Z. Luo, H. Yan, and X. Pan, 'Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques', Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.
[26] H. Yan and D. Shao, 'Enhancing Transformer Training Efficiency with Dynamic Dropout', Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.

[27] H. Yan, ‘Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing’, Optimizations in Applied Machine Learning, vol. 2, no. 1, 2022.
[28] Y. Shu, Z. Zhu, S. Kanchanakungwankul, and D. G. Truhlar, ‘Small Representative Databases for Testing and Validating Density Functionals and Other Electronic Structure Methods’, J. Phys. Chem. A, vol. 128, no. 31, pp. 6412–6422, Aug. 2024, doi: 10.1021/acs.jpca.4c03137.
[29] C. Kim, Z. Zhu, W. B. Barbazuk, R. L. Bacher, and C. D. Vulpe, ‘Time-course characterization of whole-transcriptome dynamics of HepG2/C3A spheroids and its toxicological implications’, Toxicology Letters, vol. 401, pp. 125–138, 2024.
[30] J. Shen et al., ‘Joint modeling of human cortical structure: Genetic correlation network and composite-trait genetic correlation’, NeuroImage, vol. 297, p. 120739, 2024.
[31] K. F. Faridi et al., ‘Factors associated with reporting left ventricular ejection fraction with 3D echocardiography in real‐world practice’, Echocardiography, vol. 41, no. 2, p. e15774, Feb. 2024, doi: 10.1111/echo.15774.
[32] Z. Zhu, ‘Tumor purity predicted by statistical methods’, in AIP Conference Proceedings, AIP Publishing, 2022.
[33] Z. Zhao, P. Ren, and Q. Yang, ‘Student self-management, academic achievement: Exploring the mediating role of self-efficacy and the moderating influence of gender insights from a survey conducted in 3 universities in America’, Apr. 17, 2024, arXiv: arXiv:2404.11029. doi: 10.48550/arXiv.2404.11029.
[34] Z. Zhao, P. Ren, and M. Tang, ‘Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China’, Journal of Linguistics and Education Research, vol. 5, no. 2, pp. 15–31, 2022.
[35] M. Tang, P. Ren, and Z. Zhao, ‘Bridging the gap: The role of educational technology in promoting educational equity’, The Educational Review, USA, vol. 8, no. 8, pp. 1077–1086, 2024.
[36] P. Ren, Z. Zhao, and Q. Yang, ‘Exploring the Path of Transformation and Development for Study Abroad Consultancy Firms in China’, Apr. 17, 2024, arXiv: arXiv:2404.11034. doi: 10.48550/arXiv.2404.11034.
[37] P. Ren and Z. Zhao, ‘Parental Recognition of Double Reduction Policy, Family Economic Status And Educational Anxiety: Exploring the Mediating Influence of Educational Technology Substitutive Resource’, Economics & Management Information, pp. 1–12, 2024.
[38] Z. Zhao, P. Ren, and M. Tang, ‘How Social Media as a Digital Marketing Strategy Influences Chinese Students’ Decision to Study Abroad in the United States: A Model Analysis Approach’, Journal of Linguistics and Education Research, vol. 6, no. 1, pp. 12–23, 2024.
[39] Z. Zhao and P. Ren, ‘Identifications of Active Explorers and Passive Learners Among Students: Gaussian Mixture Model-Based Approach’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, May 2025.
[40] Z. Zhao and P. Ren, ‘Prediction of Student Answer Accuracy based on Logistic Regression’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, Feb. 2025.
[41] Z. Zhao and P. Ren, ‘Prediction of Student Disciplinary Behavior through Efficient Ridge Regression’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, Mar. 2025.
[42] Z. Zhao and P. Ren, ‘Random Forest-Based Early Warning System for Student Dropout Using Behavioral Data’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, Apr. 2025.

[43] P. Ren and Z. Zhao, ‘Recognition and Detection of Student Emotional States through Bayesian Inference’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, May 2025.
[44] P. Ren and Z. Zhao, ‘Support Vector Regression-based Estimate of Student Absenteeism Rate’, Bulletin of Education and Psychology, vol. 5, no. 1, Art. no. 1, Jun. 2025.
[45] G. Zhang and T. Zhou, ‘Finite Element Model Calibration with Surrogate Model-Based Bayesian Updating: A Case Study of Motor FEM Model’, IAET, pp. 1–13, Sep. 2024, doi: 10.62836/iaet.v3i1.232.
[46] G. Zhang, W. Huang, and T. Zhou, ‘Performance Optimization Algorithm for Motor Design with Adaptive Weights Based on GNN Representation’, Electrical Science & Engineering, vol. 6, no. 1, Art. no. 1, Oct. 2024, doi: 10.30564/ese.v6i1.7532.
[47] T. Zhou, G. Zhang, and Y. Cai, ‘Unsupervised Autoencoders Combined with Multi-Model Machine Learning Fusion for Improving the Applicability of Aircraft Sensor and Engine Performance Prediction’, Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Feb. 2025, doi: 10.71070/oaml.v5i1.83.
[48] Y. Tang and C. Li, ‘Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises’, Journal of Computational Methods in Engineering Applications, pp. 1–17, 2023.
[49] C. Li and Y. Tang, ‘Emotional Value in Experiential Marketing: Driving Factors for Sales Growth–A Quantitative Study from the Eastern Coastal Region’, Economics & Management Information, pp. 1–13, 2024.
[50] C. Li and Y. Tang, ‘The Factors of Brand Reputation in Chinese Luxury Fashion Brands’, Journal of Integrated Social Sciences and Humanities, pp. 1–14, 2023.
[51] C. Y. Tang and C. Li, ‘Examining the Factors of Corporate Frauds in Chinese A-share Listed Enterprises’, OAJRC Social Science, vol. 4, no. 3, pp. 63–77, 2023.
[52] W. Huang, T. Zhou, J. Ma, and X. Chen, ‘An ensemble model based on fusion of multiple machine learning algorithms for remaining useful life prediction of lithium battery in electric vehicles’, Innovations in Applied Engineering and Technology, pp. 1–12, 2025.
[53] W. Huang and J. Ma, ‘Predictive Energy Management Strategy for Hybrid Electric Vehicles Based on Soft Actor-Critic’, Energy & System, vol. 5, no. 1, 2025.
[54] W. Huang, Y. Cai, and G. Zhang, ‘Battery Degradation Analysis through Sparse Ridge Regression’, Energy & System, vol. 4, no. 1, Art. no. 1, Dec. 2024, doi: 10.71070/es.v4i1.65.
[55] Z. Zhang, ‘RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations’, Optimizations in Applied Machine Learning, vol. 4, no. 1, 2024.
[56] Z. Zhang, K. Xu, Y. Qiao, and A. Wilson, ‘Sparse Attention Combined with RAG Technology for Financial Data Analysis’, Journal of Computer Science Research, vol. 7, no. 2, Art. no. 2, Mar. 2025, doi: 10.30564/jcsr.v7i2.8933.
[57] P.-M. Lu and Z. Zhang, ‘The Model of Food Nutrition Feature Modeling and Personalized Diet Recommendation Based on the Integration of Neural Networks and K-Means Clustering’, Journal of Computational Biology and Medicine, vol. 5, no. 1, 2025.
[58] Y. Qiao, K. Xu, Z. Zhang, and A. Wilson, ‘TrAdaBoostR2-based Domain Adaptation for Generalizable Revenue Prediction in Online Advertising Across Various Data Distributions’, Advances in Computer and Communication, vol. 6, no. 2, 2025.

[59] K. Xu, Y. Gan, and A. Wilson, ‘Stacked Generalization for Robust Prediction of Trust and Private Equity on Financial Performances’, Innovations in Applied Engineering and Technology, pp. 1–12, 2024.
[60] A. Wilson and J. Ma, ‘MDD-based Domain Adaptation Algorithm for Improving the Applicability of the Artificial Neural Network in Vehicle Insurance Claim Fraud Detection’, Optimizations in Applied Machine Learning, vol. 5, no. 1, 2025.