Overcoming Data Sparsity in Social Computing: A GMM+LLM Approach

Yulong Zhao

Department of information security, Henan University, Kaifeng, China

zhaoyulong137@outlook.com

Abstract. The rise of social media as the primary platform for high school students' interactions presents both opportunities and challenges for social computing research. While social media offers rich insights into adolescent social behavior, research often suffers from insufficient sample sizes, leading to biased portrayals of social dynamics. To address this limitation, this study proposes a novel framework that integrates Gaussian Mixture Models (GMM) with Large Language Models (LLMs) to generate linguistically detailed and context-sensitive synthetic data. Using a dataset of 15,000 student social profiles collected from 2006 to 2009, GMM clustering was applied to identify distinct user groups, followed by LLM-driven content generation tailored to each cluster through structured prompt engineering. Comparative analysis with K-means clustering demonstrates GMM' s superior ability to handle outlier-prone data, producing smoother and more representative clusters. The synthesized content successfully reflects authentic patterns of adolescent engagement across domains such as sports, music, religion, and negative topics, thereby mitigating the problem of data sparsity. This study highlights the potential of combining probabilistic modeling with advanced natural language generation to enhance the authenticity and scalability of social dynamics simulation. Future research should expand contextual features and employ robust evaluation mechanisms to further refine generative accuracy.

Keywords: Large Language Models (LLMs), Gaussian Mixture Models (GMM), CAS, Social Media Simulation, Generative Framework.

1. Introduction

With the rapid acceleration of information dissemination brought by the Internet era and the increasingly diverse lifestyles of modern high school students, social media has become the main arena of social life for this group [1]. It carries much of the unique social content created by high school students and, due to its distinctive social dynamic mechanisms, the information it stores has strong traceability. Therefore, social media serves as a valuable sample for analyzing the social lives of high school students. However, in the actual research process, observing and recording the social dynamics of a single group often faces the problem of insufficient sample size [2]. Insufficient samples weaken the accuracy of data analysis results, leading to mismatched portrayals of high school students' social roles, which may in turn result in flawed conclusions. Thus, a framework that can simulate high school social dynamics based on existing feature values can help alleviate the problem of insufficient samples.

At present, many studies have employed various methods for content synthesis. For example, Agent-based Modeling (ABM) simulates macro-level group behaviors by defining micro-level interaction rules [3]. In addition, with the recent rise of deep learning, deep generative models such as Generative Adversarial Networks (GANs) have been widely applied to synthesize social network data that conforms to real-world statistical features [4]. These studies provide important methodological foundations for constructing a framework to simulate high school students' social dynamics. However, ABM approaches struggle to capture and reproduce the rich linguistic patterns and context-dependent decision-making processes found in human social interactions. Meanwhile, GANs often face challenges such as mode collapse and training instability when generating structured text data (such as coherent social media posts).

In recent years, with the rapid development of Large Language Models (LLMs), these models have demonstrated exceptionally strong generalization and simulation capabilities in natural language

processing. Compared to ABM, which relies on preset rules to define agent behavior, LLMs provide more flexible generation and can better simulate complex behaviors in social networks, including quantifying and modeling emotional fluctuations and other irrational factors. On this basis, using LLMs to generate content as agent models can effectively improve the authenticity of simulated user behaviors in social networks.

Building on these advantages, this study aims to explore a novel approach to constructing generative agents based on LLMs to synthesize large-scale, high-quality, linguistically detailed social dynamics data of high school students, thereby providing a new solution to the problem of insufficient samples in social computing research. The research question is how to construct an adaptive generative framework for user behavior in complex social systems by combining Gaussian Mixture Models (GMM) [5] as probabilistic models with the generative capabilities of LLMs.

The remainder of this paper is structured as follows: Section 2 describe the research methodology, Section 3 provides a detailed result and a general discussion of the proposed GMM clustering + LLM generation framework, while Sections 4 conclusion.

2. Research Methodology

2.1. Data Preperation

2.1.1 About Dataset

As Table 1, the dataset was sourced from Kaggle, comprising a random sample of 15,000 high school students from 2006 to 2009 [6]. Raw data shown in Table 1 were collected by crawling dynamic content on the personal profiles of selected students from a mainstream social media application. The original researchers identified 37 high-frequency thematic keywords and quantified their occurrence frequencies to analyze the distribution of interests within the student cohort.

Table 1. Dataset structure

source	Kaggle
Samples	Social dynamics on 15000 high school students' personal homepages from 2006-2009
features	Age, Gender, Number of Friends (Network Size) and 37 High-frequency social words
Format	CSV Profile

2.1.2 Data Preprocessing and Feature Engineering

In this study, conventional data preprocessing was applied to handle outliers in which numerical variables were imputed with 0, categorical variables were filled with the mode, and the gender variable was specifically processed with "Unknown" as the replacement value. Subsequently, standardization was performed to normalize the data distribution. Additionally, the original 37 features in the dataset were complex and redundant for clustering construction. This experiment performed feature engineering to refine and supplement them.

Part 1: Feature Selection:

Five minor features with low standard deviation were pruned from the original 37 features.

Part 2: Multi-dimensional Feature Design:

Several features representing negative topics (e.g., death, drugs, alcoholism) were merged into a unified negative topic feature.

Several composite features were designed. For example, from multiple sports-related features:

1) "sports_total" feature was created to aggregate overall engagement.

2) "sports_diversity" feature was designed to measure variety.

3) "sports_top" feature was retained to capture the highest-value original sports feature in each sample.

2.2. GMM clustering + LLM generation framework

2.2.1 Introduction of GMM

The Gaussian Mixture Model (GMM) utilizes Gaussian probability functions (normal distribution curves) to precisely quantify entities by decomposing them into multiple components, each modeled based on Gaussian probability density functions. GMM is formed through a weighted sum of multiple Gaussian probability density functions. Each Gaussian component represents a specific region within the feature space. The parameters of a GMM, denoted as λ, can be expressed as:

$$\lambda = \{(p_i, \mu_i, \Sigma_i)\}_{i=1}^{M} \tag{1}$$

Where p_i represents the mixture weight, μ_i is the mean vector, and Σ_i denotes the covariance matrix. The essence of GMM lies in its ability to smoothly approximate an underlying probability distribution by combining multiple single Gaussian models.

2.2.2 Introduction of LLM

Large Language Models (LLMs) are built upon the Transformer architecture [6], which captures contextual dependencies through a self-attention mechanism. The generation process of LLMs is inherently based on sequential probability prediction. Specifically, for a given input sequence

$$X = \{x_1, x_2, \ldots, x_t\} \tag{2}$$

The model predicts the probability distribution of the next token, denoted as

$$P(x_{t+1} \mid X) \tag{3}$$

The output token is then selected via sampling strategies such as Top-p sampling.

2.2.3 How to use LLM as generator

In this experiment, structured prompt templates were employed to enforce step-by-step reasoning in the LLM. Through prompt engineering and conditional control mechanisms (e.g., temperature tuning, prefix constraints), the content generation process was guided to balance creativity and controllability [7]. This approach mitigates the risk of abrupt or inconsistent conclusions while enhancing the interpretability and traceability of the LLM's generative process. The specific sequential structure is as follows: 1) Identity anchoring, 2) Quantified Injection of Social Profile, 3) Structured Constraints for Generation Rules, 4) Language and Style Guidance, 5) Balancing Risk and Freedom, 6) Temperature Parameter and Prefix Constraints.

Each module of the prompt generator is dynamically injected based on the user cohort and date. This mechanism assists the LLM in progressively and precisely capturing the characteristics of the target group, thereby enhancing the similarity between the generated content and authentic social dynamics.

2.2.4 Framework Process

Parallel and Comparative Clustering Methodology:

Theoretically speaking, unlike traditional K-means clustering, which directly assigns each sample to a specific cluster, GMM generates a probability distribution (soft labels) indicating the likelihood of a sample belonging to each cluster. Consequently, GMM demonstrates superior performance over K-means when handling data with extreme feature values. In this experiment, GMM is introduced to generate five well-distributed and distinct clusters. The characteristic features of these clusters will be utilized to develop a personalized caption generator tailored to the corresponding groups. As what is shown in Fig. 1, in this study, a comparative analysis was conducted between the K-means and GMM clustering algorithms, designated as the control and experimental groups respectively, to evaluate their comparative efficacy on the given dataset within the proposed framework.

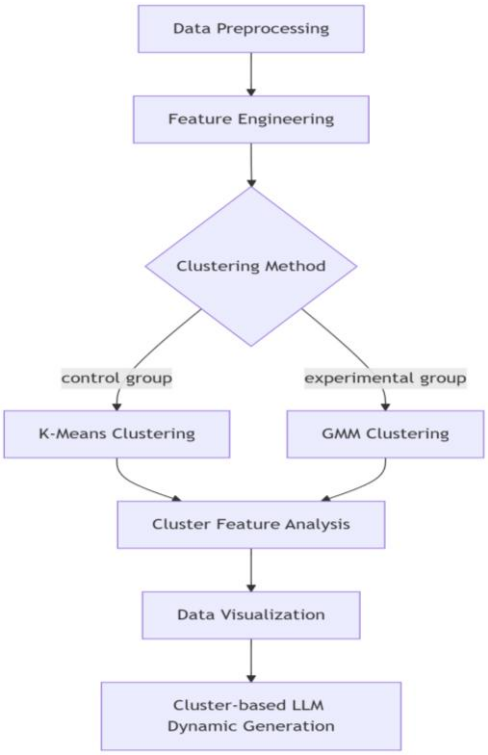

Figure 1. The whole framework process (Picture credit: Original)

Cluster Feature Analysis:

Centering on the standard deviation of feature values at the cluster centroids, this study conducted a multi-dimensional analysis of the clustering results, evaluating the significance of different feature metrics and the inter-feature correlations within clusters.

LLM Dynamic Generation (Fig. 2):

Prompt Design

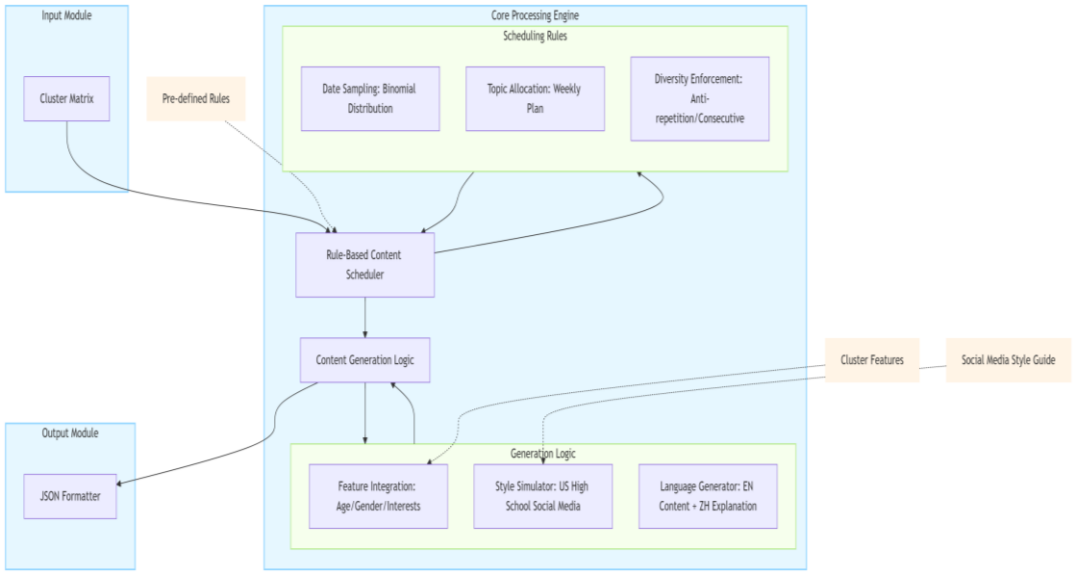

Figure 2. Generation Process (Picture credit: Original)

A Dynamic Prompt Generator Based on User Cohort Profiling:

This prompt generator dynamically adjusts itself according to the distinct feature means of different clusters. For example, based on the "sports_total" and "sports_diversity" features, it dynamically selects tailored sports descriptions "sports_des" for specific user groups.

This design enhances the LLM's ability to sensitively capture subtle distinctions between different user cohorts.

Hyperparameter Configuration

The API configuration for the invoked LLM (DeepSeek V2) provides several hyperparameters to modulate the model's output:

"temperature": Controls the randomness and creativity of the generated text.

"top_p": The model samples only from the most probable tokens whose cumulative probability mass just exceeds the "top_p" threshold.

"frequency_penalty": Reduces the likelihood of the model repetitively using the same vocabulary.

"presence_penalty": Discourages the model from repeatedly addressing the same topics.

In this study, iterative controlled experiments were conducted to optimize these hyperparameters, aiming to approach a balance between the diversity and authenticity of the generated content.

3. Result and Discussion

3.1. Clustering results visualization

Figure 3. Comparison of clustering performance between GMM and K-means algorithms (Picture credit: Original)

Fig. 3 is a pair of pictures which show the different results between K-means clustering algorithm and GMM clustering algorithm. The K-means clustering algorithm exhibits a cluster centroid that significantly deviates from most samples, whereas all centroids in GMM are positioned within dense sample regions. This visual representation demonstrates GMM's superior ability to capture authentic data distribution patterns compared to the distance-based partitioning of K-means.

The K-means algorithm demonstrates sensitivity to outliers due to its underlying assumption of uniformly sized, rigid spherical clusters. Distant data points significantly distort centroid calculations during iteration. In contrast, the Gaussian Mixture Model (GMM) operates under more flexible assumptions, accommodating ellipsoidal clusters of varying sizes and orientations through covariance matrices. This probabilistic approach endows GMM with enhanced robustness when handling noisy or outlier-prone datasets compared to the deterministic partitioning of K-means.

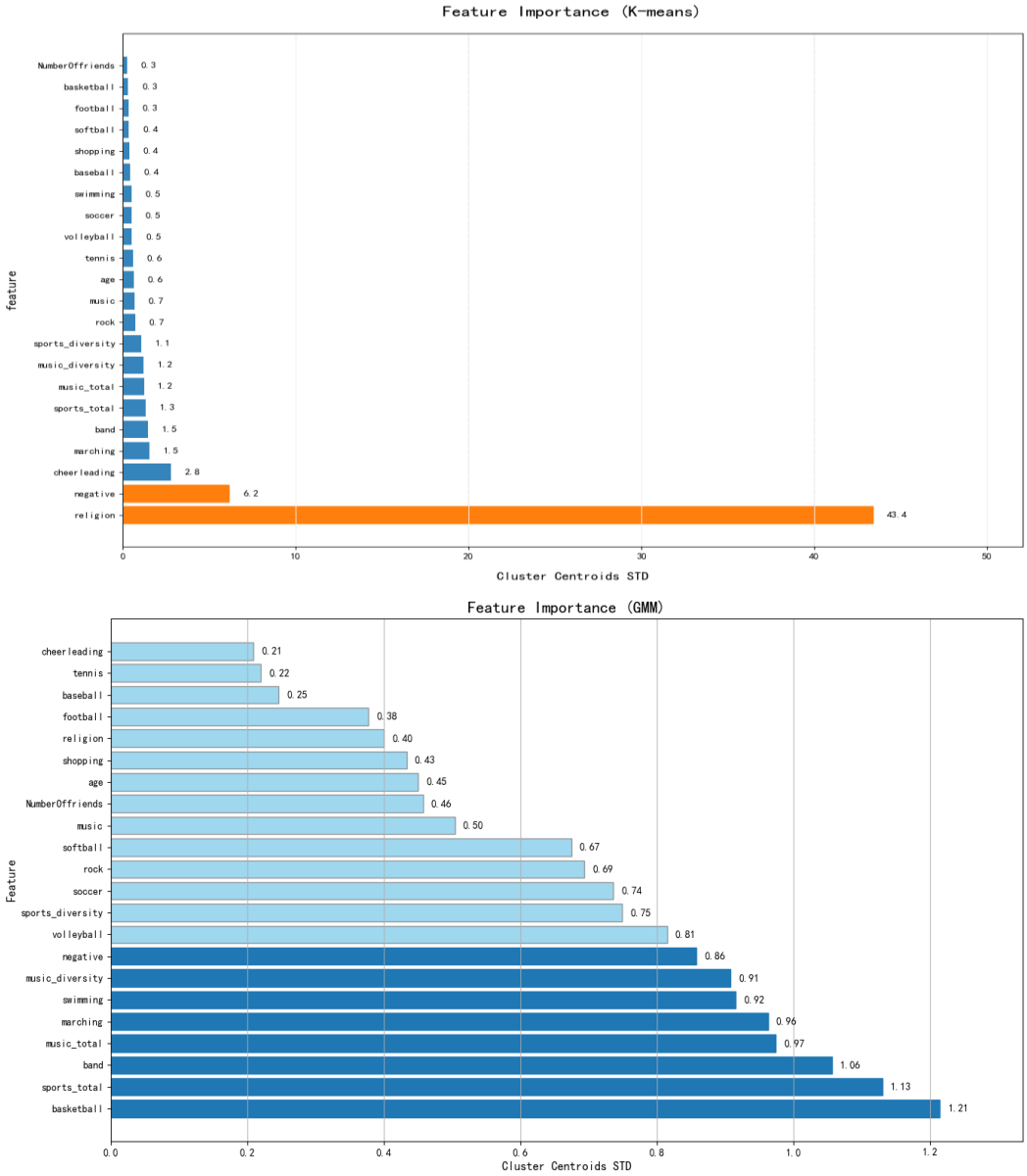

Figure 4. Feature Importance Distribution (Picture credit: Original)

Fig. 4 demonstrates the importance of each feature about how much they influence the clustering result. In this study, the importance of a feature is interpreted as its relative weight in distinguishing between different clusters. Thus, the standard deviation of each feature's value across the cluster centroids was employed to quantify the degree of variation for the corresponding feature among the clusters. A higher degree of variation indicates that the feature plays a more significant role in differentiating the clusters and therefore possesses greater importance.

A pronounced distinction is observed between the clustering results: the GMM-derived bar chart exhibits a smoothly varying gradient of feature importance, whereas the K-means result demonstrates an extreme outlier where the standard deviation of the religion feature drastically exceeds all other values, effectively functioning as the decisive factor in cluster separation. This pattern underscores the vulnerability of K-means to extreme sample values, further validating its limitations in handling datasets with outlier-prone distributions.

3.2. Cluster Feature Analysis

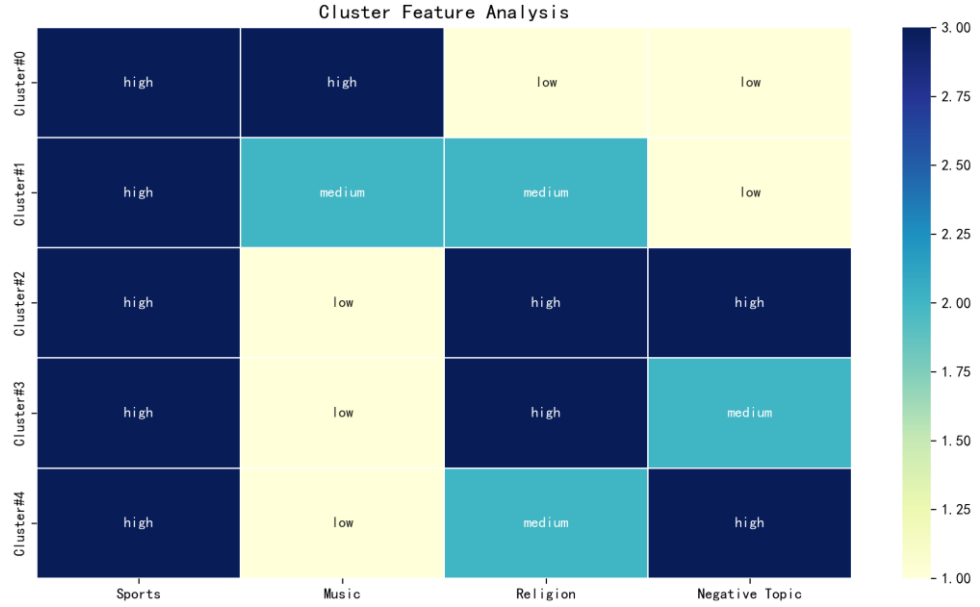

Figure 5. Heatmap of Cluster Features (Picture credit: Original)

This Heatmap in Fig. 5 illustrates distinct level of interest from five respective clusters across four key thematic domains: Sports, Music, Religion, and Negative Topics. The intensity of each feature, quantified as High, Medium, or Low, is encoded by color saturation.

Based on their different behavioral profiles, this paper formulates a general personality framework for each cluster to demonstrate the result of the clustering algorithm.

Cluster #0: The "Sports & Music Enthusiasts": This group exhibits dominant engagement in Sports and Music with negligible interest in Religion or Negative Topics. They represent a mainstream, well-adjusted demographic.

Cluster #1: The "Moderately Engaged" Cluster: Characterized by a high interest in Sports and moderate engagement in both Music and Religion, this cluster shows a balanced profile with low negativity.

Cluster #2 & #4: The "Religion-Focused" Groups: Both clusters show high engagement with Religion and Negative Topics, coupled with low interest in Music. Cluster #2 exhibits a particularly high level of negative expression, suggesting a distinct subgroup within this demographic. Cluster #4 shows a slightly more moderate profile.

Cluster #3: The "Sports-Focused Religious" Cluster: This group shares the high Sports engagement of Clusters #0 and #1 but combines it with a strong focus on Religion and a medium level of Negative Topics, presenting a unique hybrid profile.

Cluster Characteristics Analysis:

Cluster #0 and #1 exhibit stronger emphasis on sports and music.

Cluster #2 and #3 demonstrate significant engagement in sports and religion.

Cluster #4 is characterized by sports diversity and potential negative themes.

1) User groups with higher values in music-related features generally show elevated values in negative topics, indicating a weak positive correlation between music engagement and negative expression. 2) A divergent correlation exists between religious feature values and negative topics:

Cluster #2 (highest religion value) shows the highest negative topic value, while Cluster #3 (lowest religion value) exhibits the lowest. Groups with moderate religious engagement display relatively lower negative topic values.

3)A noticeable positive correlation is observed between sports engagement and negative topics: Cluster #2, with the highest sports feature value, also displays the highest negative topic value.

4) Cluster #3, with low values across all features, shows the lowest negative topic value. Conversely, groups with higher overall feature values tend to exhibit higher negative topic values.

5) Sports feature values and religious feature values show a weak negative correlation.

3.3. Discussion

Through repeated hyperparameter tuning and result validation, this study demonstrates that the GMM+LLM framework successfully generates vivid and diverse content that aligns with both general high school social media conventions and the specific characteristics of each identified user cluster. Monthly statistical analysis confirms that the generated content tags consistently reflect the expected feature values of their respective clusters, thereby addressing the core research problem outlined in the introduction. However, the generated content remains primarily constrained to the feature dimensions present in the original dataset. To achieve a more comprehensive simulation of high school social dynamics, future work should expand the dataset dimensions to incorporate additional contextual factors such as academic activities, family relationships, friendship networks, and intimate relationships, which would enable the construction of more precise user profiles. Furthermore, to objectively evaluate content quality, subsequent studies should implement robust evaluation mechanisms including Siamese networks [8, 9], generative adversarial networks (GANs) [10], or structured questionnaires to quantitatively measure the quality gap between generated content and authentic social media dynamics.

4. Conclusion

This paper presented a framework that uses GMM and LLM to overcome the lack of data in social computing studies. The results show that GMM performs better than K-means when clustering noisy data, and the LLM can produce texts that match the main features of each cluster. The framework creates content that is both diverse and close to real social media use among students. Still, the work has limits because the dataset only covers certain features. More factors, such as school activities or family ties, should be included in the future to build fuller profiles. Evaluation methods like questionnaires or advanced models could also help check the quality of generated content. Overall, this research offers a practical way to reduce data sparsity and improve simulations of student social life.

References

[1] Smith A, Anderson M, Jiang J. Social media use in 2021: A deep dive into the adolescent experience. J Adolesc Health. 2021; 68 (2): 145-52.

[2] Chen L, Wang H. Data sparsity and bias in computational social science: A review of challenges and solutions. IEEE Trans Comput Soc Syst. 2022; 9 (4): 1120-32.

[3] Squazzoni F, Polhill GJ, Edmonds B, et al. Computational models that matter during a global pandemic: The case of COVID-19. ACM Comput Surv. 2020; 53 (6): 1-37.

[4] Li Y, Wang D, Sun Y. SparseGAN: Synthesizing social network data with generative adversarial networks for research on rare populations. IEEE Trans Knowl Data Eng. 2021; 34 (8): 3892-905.

[5] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process. 1995; 3 (1): 72-83. doi: 10.1109/89.365379.

[6] Zabihullah18. Students' social network profile clustering [dataset]. Kaggle. 2020. Available from: https://www.kaggle.com/datasets/zabihullah18/students-social-network-profile-clusterin.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017; 30: 5998-6008.

[8] He A, Luo C, Tian X, Zeng W. A twofold siamese network for real-time object tracking. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2018. p. 4834-43.

[9] Melekhov I, Kannala J, Rahtu E. Siamese network features for image matching. In: Proc 23rd Int Conf Pattern Recognit (ICPR). IEEE; 2016. p. 378-83.

[10] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Process Mag. 2018; 35 (1): 53-65.