Evaluating Synthetic Data Generation Methods for Anomalous Channel Detection in Sparse-Label Environments

Ridwan Amure[1] and Nitin Agarwal[1,2]

[1] COSMOS Research Center, University of Arkansas at Little Rock, Little Rock, AR 72204, USA
raamure@ualr.edu, nxagarwal@ualr.edu
[2] International Computer Science Institute, University of California, Berkeley, CA 94720, USA

Abstract. Detecting anomalous behavior on YouTube is hindered by the limited availability of labeled data. This study examines the application of synthetic data generation to enhance anomaly detection in data-scarce environments. We analyze a dataset of 97 YouTube channels—7 of which were suspended—covering over 640,000 videos and 123 million comments. Structural and engagement features were extracted to train four synthetic data generators: CTGAN, TVAE, Gaussian Copula, and HMASynthesizer. Classifiers, including Logistic Regression, Random Forest, and SVMs, were trained exclusively on the synthetic data and evaluated on real, withheld anomalies. Results show that models trained on CTGAN- and TVAE-generated data achieved F1 scores up to 0.875, demonstrating strong detection capability. These findings highlight the potential of synthetic augmentation as a practical solution to the labeled data bottleneck in social media anomaly detection.

Keywords: Synthetic data, anomaly detection, YouTube, social media analytics, CTGAN, TVAE, low-resource learning

1 Introduction

YouTube, the world's largest video-sharing platform, has evolved into a central space for information exchange, entertainment, and public discourse. With billions of users and hundreds of hours of content uploaded every minute, the platform exerts a profound influence on global narratives. From shaping political opinions to promoting social movements, YouTube plays a key role in how individuals and communities perceive the world. However, its openness and scale also make it a fertile ground for inauthentic and manipulative behavior.

The challenge of detecting such behavior, especially at the channel level, is increasingly urgent. Suspended or malicious channels often exhibit subtle engagement strategies that exploit platform dynamics, such as comment coordination, artificial amplification, and stealthy dissemination of controversial content. In regions with high geopolitical tension—such as the Indo-Pacific or East

Asia—these activities can be particularly consequential, influencing public sentiment, fueling misinformation, or exacerbating conflict [13]

Recent work by Amure and Agarwal [3] highlighted a critical bottleneck in anomaly detection efforts on YouTube: the scarcity of labeled anomalous data. In their study of anomalous channel behavior, they highlighted the challenge of obtaining sufficient verified cases of suspension or malicious coordination to support supervised learning. This challenge is further exacerbated by the fact that anomalous behavior is often rare, temporally bursty, and contextually specific, making generalization difficult. As a result, conventional approaches—especially those requiring large labeled datasets—struggle to scale.

To address this problem, researchers have begun exploring synthetic data generation as a promising strategy. Synthetic data can augment small datasets, enabling models to learn more effective decision boundaries and enhance their ability to detect rare anomalies. In domains such as cybersecurity [5], fraud detection [17], and medical diagnostics [4], synthetic sampling techniques—ranging from copula models to GANs and variational autoencoders—have shown measurable improvements in detection tasks under data scarcity.

In this study, we apply this principle to the problem of detecting anomalies in YouTube channels. Our dataset comprises structural and engagement-based features from 97 YouTube channels, including seven suspended channels identified as anomalous. Building on the experimental framework outlined in Amure and Agarwal [3], we simulate low-label scenarios by withholding a subset of these anomalies and investigate whether synthetic data generation can enhance model performance in identifying withheld anomalous cases.

We investigated the use of synthetic data generation to support anomaly detection in scenarios with limited labeled data. Our study presents a reproducible framework that identifies and removes known anomalies, generating synthetic training data from the remaining real samples. We compared four generation methods—Gaussian Copula Synthesizer [12], Conditional Tabular Generative Adversarial Network (CTGAN) [16], Tabular Variational Autoencoder (TVAE) [16], and Hierarchical Multi-Table Modeling with HMASynthesizer [12]—based on their suitability for generating tabular data suitable for classification. Through repeated experiments, we provide early insights into the reliability and usefulness of synthetic augmentation in detecting real-world anomalies, offering practical guidance for researchers working in low-resource settings.

2 Related Work

Detecting anomalous behavior on YouTube has attracted increasing attention as the platform continues to shape public discourse and influence global narratives. Researchers have approached this challenge from different angles, focusing on both engagement metrics and the structural behavior of users.

One prominent line of work has focused on coordinated inauthentic behavior, where users or groups artificially inflate engagement to manipulate visibility or trust. For example, Kirdemir et al. [10] observed that channels engaged in such

activity often display sudden bursts in engagement, with fewer but more intense peaks. Similarly, Hussain et al. [7] and Adeliyi et al. [1] explored patterns of inorganic engagement using unsupervised techniques, revealing how channels can be systematically used to spread misinformation or disrupt authentic discourse.

Other studies have focused on the behavior of commenters and the structure of the network. Shajari et al. [14] proposed analyzing co-commenter networks—where connections are formed between users commenting on the same videos—to uncover tightly connected cliques indicative of coordination. They employed techniques such as Principal Component Analysis (PCA) and clustering to isolate suspicious activity. Building on this, Shajari et al. [15] introduced a normalized anomaly score based on 20 structural features of the co-commenter network, using Kernel Density Estimation (KDE) and Gaussian Mixture Models (GMM) to differentiate anomalous channels. Their work was particularly focused on channels operating in the Indo-Pacific region, a context marked by geopolitical tensions and coordinated influence efforts.

Beyond political or manipulative content, safety concerns have also been studied. Kaushal et al. [8] utilized convolutional neural networks (CNNs) to identify unsafe video content targeted at children and found that promoters of unsafe content are often structurally close to legitimate ones in the engagement network. Papadamou et al. [11] echoed these concerns, demonstrating how inappropriate content can bypass YouTube's detection systems by mimicking benign videos.

Building on the role of network structures, Akinnubi et al. [2] emphasized the utility of graph-based models for uncovering unscrupulous actors and engagement manipulation across platforms. These findings reinforce the value of co-commenter and channel-level networks in capturing the dynamics of abnormal behavior.

User motivations and interaction patterns have also been taken into consideration. Galeano et al. [6] emphasized that commenters play an active role in shaping public perception and amplifying narratives—especially in disinformation contexts. Khan [9] identified social interaction as a core driver behind commenting behavior, suggesting that genuine and inauthentic engagements can sometimes be behaviorally similar, complicating detection.

Despite these advancements, one persistent challenge in this area remains the lack of labeled data for known anomalies. As Amure and Agarwal [3] emphasized, it is rare to have confirmed information about which channels were suspended due to policy violations. This scarcity of labeled examples limits the effectiveness of supervised learning models and motivates the need for alternative strategies.

In response, our study takes a novel approach by exploring the use of *synthetic data generation* to support anomaly detection when labeled data is limited. Synthetic data has shown promise in fields like fraud detection [5], cybersecurity, and medical imaging [4], yet its application in the social media anomaly space remains underexplored. We aim to fill this gap by systematically evaluating whether synthetic samples can provide meaningful support to classifiers

trained to detect anomalous YouTube channels—particularly in realistic, low-label scenarios.

3 Data

To support our investigation, we curated a comprehensive dataset using a specialized data extraction tool developed by Kready et al. [**?**], which interfaces with the official YouTube Data API [**?**]. This tool enabled us to gather detailed, structured information about video uploads, comment activity, and user interactions across a select group of YouTube channels.

Our dataset comprises 97 YouTube channels, seven of which were suspended by the platform for violating community guidelines—thus providing labeled examples of anomalous behavior. Across these channels, we collected metadata on over 640,000 videos, more than 123 million comments, and interactions involving upwards of 12 million unique commenters.

The selection of channels and associated content was guided by thematic relevance to the Indo-Pacific region—a context often tied to geopolitical discourse, misinformation, and digitally coordinated campaigns. Keywords used for channel discovery included phrases such as *"Komunis Cina — China pengaruh Indonesia"*, *"Muhammadiyah Cina — China — Tiongkok — Tionghoa"*, and *"Kejam Uighur — Uyghur"*, which were identified through coverage analysis and refined through iterative review to enhance inclusiveness and relevance.

To model social interactions, we constructed a co-commenter network by linking commenter nodes based on shared activity. Specifically, an undirected edge was formed between two commenters if they had posted at least five comments on videos within the same channel—a threshold previously validated by Shajari et al. [13] as sufficient to suggest coordinated or recurrent engagement. Each commenter was also linked to the corresponding channel node, creating a two-layer structure of user-user and user-channel relationships.

For each channel's co-commenter network, we extracted a total of twenty structural features, denoted as F_s, as introduced by Shajari et al. [13]. These features capture various graph-theoretic properties of commenter interactions—such as node count, edge density, clustering coefficients, and clique-based metrics—that help characterize the underlying engagement structure. In addition to these, we incorporated a complementary set of engagement-based features, denoted as F_e, adapted from Kirdemir et al. [10]. These include ratios and trends based on views, comments, and subscriber counts, which provide insight into anomalous content performance and growth dynamics. For further technical details on the design and interpretation of these feature sets, we refer interested readers to the respective studies [10, 13].

4 Method

The primary objective of this study is to assess the effectiveness of synthetic data generation techniques in enhancing anomaly detection in data-scarce environ-

ments. We propose a controlled evaluation framework that simulates a low-label setting commonly encountered in real-world social media analytics.

We began by extracting feature representations for each channel, comprising both structural features (F_s) and engagement features (F_e), as described earlier. Using this dataset, we applied four different synthetic data generation methods to produce additional samples for model training. These generators were trained on a subset of real channels with known structural and engagement characteristics.

To assess their utility, we conducted the following experimental procedure:

1. We randomly selected and removed four known anomalous channels (i.e., suspended channels) from the dataset. These channels, along with a random sample of normal channels, formed the test set.
2. The remaining data was used to fit each synthetic data generator, which then produced synthetic samples matching the structure of the original dataset.
3. We trained five standard classifiers—*Logistic Regression*, *Random Forest*, and three variants of Support Vector Machines with polynomial, sigmoid, and radial basis function (RBF) kernels—using only the generated synthetic data.
4. Each trained model was evaluated on the withheld test set, which contained real (non-synthetic) samples, including the known anomalies.

This process was repeated across ten randomized trials to ensure robustness. In the following subsections, we describe the four synthetic data generators used in our experiments: Gaussian Copula, Conditional Tabular GAN (CTGAN), Tabular Variational Autoencoder (TVAE), and the Hierarchical Multi-table Synthesizer (HMASynthesizer).

4.1 Synthetic Data Generators

This study examines four prominent synthetic data generation techniques tailored for structured tabular data. These methods are designed to replicate the statistical properties and interdependencies of real-world observations, enabling the creation of artificial datasets that reflect the complexity of structural and engagement patterns found in YouTube channel behavior.

The chosen generators represent a diverse range of modeling strategies. Each generator is described below.

Gaussian Copula Synthesizer: The Gaussian Copula Synthesizer [12] generates synthetic data by modeling the dependency structure among variables using a copula, allowing for the separation of marginal distributions from their joint dependence structure.

Let $\mathbf{X} = [X_1, X_2, \ldots, X_d]$ be a d-dimensional random vector. Each X_i is transformed into a uniform variable $U_i = F_i(X_i)$, where F_i is the empirical cumulative distribution function (CDF). A Gaussian copula C_Σ models the dependency:

$$C_\Sigma(u_1, \ldots, u_d) = \Phi_\Sigma\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

where Φ^{-1} is the inverse CDF of the standard normal distribution and Φ_Σ is the joint CDF of a multivariate normal with covariance matrix Σ.

New samples are drawn by sampling from $\mathcal{N}(0, \Sigma)$, converting them to uniform marginals, and transforming back using the inverse of the empirical CDFs.

Tabular Variational Autoencoder (TVAE): TVAE [16] is a variational autoencoder (VAE) adapted to tabular data, modeling its latent structure and generating samples via reparameterization.

The model optimizes the evidence lower bound (ELBO):

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \| p(z)),$$

where $q_\phi(z|x)$ and $p_\theta(x|z)$ are the encoder and decoder, respectively, and $p(z)$ is a standard normal prior. After training, samples are generated by drawing $z \sim \mathcal{N}(0, I)$ and decoding to the feature space.

Hierarchical Multi-table Synthesizer (HMASynthesizer): The HMASynthesizer [12] is designed for multi-table (relational) data synthesis. It models the joint distribution over a set of tables by learning their topological dependencies and synthesizing them hierarchically.

If $T_1, T_2, ..., T_n$ are relational tables, then:

$$P(T_1, T_2, ..., T_n) = \prod_{i=1}^{n} P(T_i \mid \text{parents}(T_i)).$$

For our single-table YouTube data, HMASynthesizer operates in "flattened" mode, learning attribute-level dependencies across all rows and columns using Bayesian networks or copula-based models.

4.2 Evaluation

We assessed model performance using the F1 score, which balances precision and recall to reflect a classifier's ability to correctly identify anomalies without over-predicting them. All evaluations were conducted on a real test set that included the four withheld anomalous channels alongside a sample of normal channels. To ensure consistency, we averaged results over ten independent trials.

The F1 score is defined as:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \quad \mathrm{Precision} = \frac{TP}{TP + FP}, \quad \mathrm{Recall} = \frac{TP}{TP + FN}$$

This metric is critical in imbalanced settings, where accuracy alone may be misleading. F1 provides a more reliable measure of how well models identify rare anomalous cases.

As baselines, we include a random classifier (F1 score ≈ 0) and a real-only classifier trained on the same subset of real channels used to fit the data generators (excluding the held-out anomalies) (F1 score ≈ 0.2). These serve as reference points: the random classifier represents a lower bound, while the real-only classifier offers a benchmark for performance achievable using actual labeled data without synthetic augmentation.

5 Result and Discussion

Our evaluation reveals several important patterns regarding the interplay between data synthesizers and downstream classifiers, as shown in Figure 1. Among the combinations tested, models trained on synthetic data generated by CTGAN and TVAE exhibited the strongest performance, particularly when paired with Random Forest classifiers. For instance, CTGAN combined with Random Forest achieved an F1 score of 0.875, outperforming all other configurations. This suggests that CTGAN not only models the data distribution effectively but also preserves decision-relevant structures needed for accurate anomaly classification. A similar trend was observed with TVAE, which achieved an F1 score of 0.727 under Random Forest, further demonstrating the potential of deep generative models for augmenting detection in data-scarce environments.

Models trained on data generated by the Gaussian Copula Synthesizer exhibited relatively stable performance across classifiers, although they were slightly less expressive than CTGAN and TVAE. While Random Forest achieved an F1 score of 0.667 under Gaussian Copula, SVC with an RBF kernel reached 0.750, suggesting that margin-based classifiers better leverage the statistical structure captured by the copula than tree-based ones.

The HMASynthesizer yielded more mixed results. While some configurations achieved moderate F1 scores, others performed poorly—for example, Random Forest and SVC-Sigmoid achieved F1 scores of only 0.421 and 0.428, respectively. These inconsistencies suggest that while HMASynthesizer may model hierarchical dependencies effectively, it may fall short in preserving localized feature patterns essential for identifying sparse anomalies.

Across all generators, Random Forest consistently delivered strong results, highlighting its robustness in handling synthetic feature variability and nonlinear interactions. Logistic Regression and SVC with polynomial kernels performed moderately under Gaussian Copula and TVAE but struggled when trained on CTGAN and HMA-generated data. In contrast, SVC with a sigmoid kernel underperformed across all settings, with F1 scores rarely exceeding 0.5—likely due to its limited flexibility in capturing complex decision boundaries.

A key insight from our experiments is the sensitivity of classification outcomes to both the choice of synthetic generator and the downstream model. While CTGAN and TVAE offer strong generative capabilities, their success depends on the classifier's ability to extract relevant patterns from synthetic data. More broadly, our results affirm the viability of synthetic data as a solution to the label scarcity problem in social media anomaly detection. They also emphasize the

importance of focusing on threshold-sensitive metrics, such as the F1 score, which better reflect operational effectiveness in imbalanced and high-stakes settings than ranking-based alternatives.

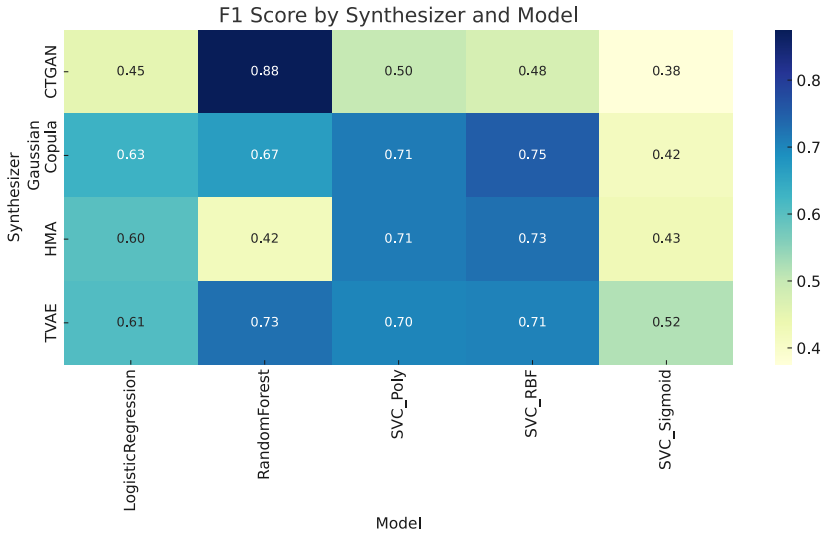

Fig. 1. F1 Score of classifiers trained on synthetic data generated by four different SDV synthesizers. Each cell represents the F1 score of a model trained exclusively on synthetic samples and evaluated on real hold-out data containing both anomalous and non-anomalous channels.

6 Conclusion, Limitations, and Future Work

This study demonstrates that synthetic data generation can effectively address the lack of labeled anomalies in social media research. By training classifiers solely on synthetic data, we demonstrated that methods such as CTGAN and TVAE produce samples that support strong anomaly detection performance—especially when paired with Random Forest classifiers. Our aim was not to identify the best model, but to highlight how synthetic augmentation can alleviate the low-data bottleneck.

We acknowledge that further model tuning could enhance the results and that our current setup does not fully explore multimodal inputs, temporal features, or distributional shifts between real and synthetic data. Future work will extend this framework to larger datasets, improve domain adaptation, and explore more advanced representations for capturing social behavior.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Adeliyi, O., Solaiman, I., Shajari, S., Onyepunuka, U., Agarwal, N.: Detecting and characterizing inorganic user engagement on youtube. In: Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media (2024). DOI 10.36190/2024.01
2. Akinnubi, A., Alassad, M., Agarwal, N., Amure, R.: Identifying contextualized focal structures in multisource social networks by leveraging knowledge graphs. In: International Conference on Complex Networks and Their Applications, pp. 15–27. Springer (2023)
3. Amure, R., Agarwal, N.: Anomalous channel detection for youtube through label propagation. In: International Conference on Complex Networks and Their Applications, pp. 271–281. Springer (2024)
4. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology **65**(5), 545–563 (2021). DOI 10.1111/1754-9485.13261
5. Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F.: Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences **479**, 448–455 (2019). DOI 10.1016/j.ins.2018.02.060
6. Galeano, K., Galeano, R., Agarwal, N.: An evolving (dis) information environment–how an engaging audience can spread narratives and shape perception: A trident juncture 2018 case study. Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities pp. 253–265 (2020)
7. Hussain, M.N., Tokdemir, S., Agarwal, N., Al-Khateeb, S.: Analyzing disinformation and crowd manipulation tactics on youtube. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1092–1095. IEEE (2018)
8. Kaushal, R., Saha, S., Bajaj, P., Kumaraguru, P.: KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. In: 2016 14th Annual Conference on Privacy, Security and Trust (PST), pp. 157–164. IEEE (2016). URL https://ieeexplore.ieee.org/abstract/document/7906950/

9. Khan, M.L.: Social media engagement: What motivates user participation and consumption on youtube? Computers in human behavior **66**, 236–247 (2017)
10. Kirdemir, B., Adeliyi, O., Agarwal, N.: Towards characterizing coordinated inauthentic behaviors on youtube. In: ROMCIR@ ECIR, pp. 100–116 (2022)
11. Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., Sirivianos, M.: Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children. In: Proceedings of the international AAAI conference on web and social media, vol. 14, pp. 522–533 (2020)
12. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. IEEE (2016). DOI 10.1109/DSAA.2016.49
13. Shajari, S., Amure, R., Agarwal, N.: Analyzing anomalous engagement and commenter behavior on youtube. In: AMCIS 2024 Proceedings (2024). URL https://aisel.aisnet.org/amcis2024/social_comp/social_comput/6/
14. Shajari, S., Amure, R., Agarwal, N.: Detecting and measuring anomalous behaviors on youtube. In: Social Networks Analysis and Mining. Springer Nature Switzerland (2025)
15. Shajari, S., Amure, R., Agarwal, N.: Navigating the anomalies: A comprehensive analysis of youtube channel behavior. In: Social Networks Analysis and Mining. Springer Nature Switzerland (2025)
16. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. In: Advances in Neural Information Processing Systems, vol. 32, pp. 7333–7343 (2019)
17. Zhang, Z., Li, Y., Ren, Y., Wang, D.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. Proceedings of the 2019 World Wide Web Conference p. 187–196 (2020). DOI 10.1145/3308558.3313487