Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 14(2): 760-773 (2026)

Evaluating Minority-Class Detection in Imbalanced Hate Speech Data via GAN-Based Synthetic Text Generation and LSHADE Hyperparameter Optimization

Bekir DEMİRAĞ[1*] Yusuf SÖNMEZ[2] Hamdi Tolga KAHRAMAN[3]

[1]*Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Turkey*

[2]*Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Turkey*

[3]*Karadeniz Technical University, Faculty of Engineering, Department of Industry Engineering, Trabzon, Turkey*

Article Info

Research article
Received: 02/05/2026
Revision: 26/05/2026
Accepted: 01/06/2026

Keywords

Hate Speech Detection, Generative Adversarial Networks, Metaheuristic Optimization, Class Imbalance, LSHADE, SMOTE, ADASYN

Makale Bilgisi

Araştırma makalesi
Başvuru: 02/05/2026
Düzeltme: 26/05/2026
Kabul: 01/06/2026

Anahtar Kelimeler

Nefret Söylemi Tespiti, Üretici Çekişmeli Ağlar, Meta-Sezgisel Optimizasyon, Sınıf Dengesizliği, LSHADE, SMOTE, ADASYN

Graphical/Tabular Abstract (Grafik Özet)

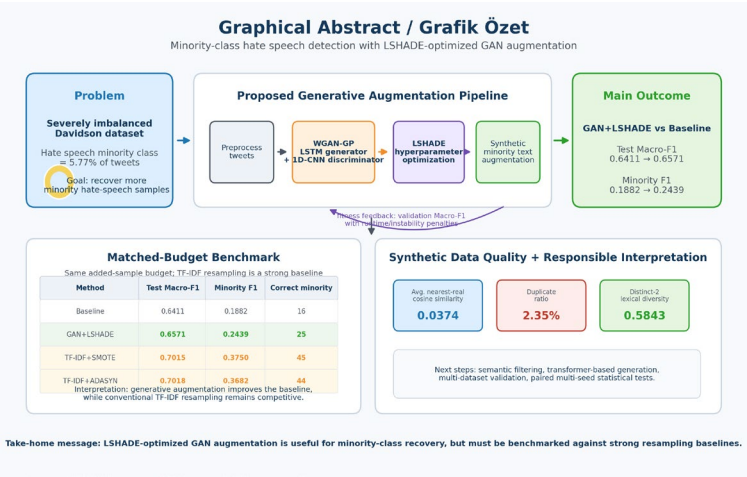

Figure A*: LSHADE-optimized GAN-based synthetic text generation is evaluated for imbalanced hate speech detection and compared with SMOTE/ADASYN. /* ***Şekil A:*** *Dengesiz nefret söylemi tespitinde LSHADE optimize GAN tabanlı sentetik veri üretimi değerlendirilmiş ve SMOTE/ADASYN ile karşılaştırılmıştır.*

Highlights (Önemli noktalar)

- *LSHADE-optimized WGAN-GP improves minority-class hate speech detection. /LSHADE optimize WGAN-GP, azınlık sınıfı nefret söylemi tespitini iyileştirir.*
- *GAN+LSHADE increases correctly detected hate speech samples from 16 to 25. /GAN+LSHADE doğru tespit sayısını 16'dan 25'e çıkarmıştır.*
- *TF-IDF-based SMOTE/ADASYN remain strong competitors. / TF-IDF tabanlı SMOTE/ADASYN güçlü rakiplerdir.*
- *Synthetic texts show low duplication and limited memorization. / Sentetik metinlerde düşük tekrar ve sınırlı ezberleme gözlenmiştir.*

Aim (Amaç): *This study evaluates LSHADE-optimized GAN-based synthetic text generation for improving minority-class hate speech detection in imbalanced datasets. / Bu çalışma, dengesiz nefret söylemi veri kümelerinde LSHADE optimize GAN tabanlı sentetik metin üretiminin etkisini değerlendirmektedir.*

Originality (Özgünlük): *The study integrates GAN-based text generation with LSHADE optimization and compares it with TF-IDF-based resampling methods under equal sample budgets. / Çalışma, GAN tabanlı metin üretimini LSHADE optimizasyonu ile birleştirerek TF-IDF tabanlı yeniden örnekleme yöntemleriyle karşılaştırmaktadır.*

Results (Bulgular): *GAN+LSHADE improves baseline Macro-F1 and minority-class F1 scores, but SMOTE and ADASYN achieve higher performance in matched-budget comparisons. / GAN+LSHADE temel modele göre Macro-F1 ve azınlık sınıfı F1 skorlarını artırmış; ancak SMOTE ve ADASYN daha yüksek performans göstermiştir.*

Conclusion (Sonuç): *LSHADE-optimized GAN augmentation is useful for minority-class improvement, but it is not universally superior to traditional resampling methods. / GAN+LSHADE, azınlık sınıfı tespitinde yararlı bir artırma stratejisidir; ancak geleneksel yeniden örnekleme yöntemlerine her durumda üstün değildir.*

*Corresponding author, e-mail: 23833301013@gazi.edu.tr
DOI: 10.29109/gujsc.1942853

GU J Sci, Part C, 14(2): 760-773 (2026)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Evaluating Minority-Class Detection in Imbalanced Hate Speech Data via GAN-Based Synthetic Text Generation and LSHADE Hyperparameter Optimization

Bekir DEMİRAĞ[1*] Yusuf SÖNMEZ[2] Hamdi Tolga KAHRAMAN[3]

[1]*Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Turkey*

[2]*Gazi University, Faculty of Technology, Department of Computer Engineering, Ankara, Turkey*

[3]*Karadeniz Technical University, Faculty of Engineering, Department of Industry Engineering, Trabzon, Turkey*

Article Info

Research article
Received: 02/05/2026
Revision: 26/05/2026
Accepted: 01/06/2026

Keywords

Hate Speech Detection, Generative Adversarial Networks, Metaheuristic Optimization, Class Imbalance, LSHADE, SMOTE, ADASYN

Abstract

The rapid spread of hate speech on social media makes automatic detection systems indispensable; however, severe class imbalance in real-world datasets substantially limits the effectiveness of deep learning models. This study proposes an end-to-end framework that integrates a Long Short-Term Memory (LSTM) generator and a one-dimensional Convolutional Neural Network (1D-CNN) discriminator within a WGAN-GP architecture to address the imbalance problem in the Davidson hate speech dataset, where the minority class represents only 5.77% of all samples. To mitigate the training instability of standard GANs and the high computational cost of conventional grid search, the LSHADE metaheuristic is employed to autonomously optimize the main GAN hyperparameters. In response to the review comments, the revised manuscript adds a clearer justification for choosing LSHADE, clarifies how gradient flow is handled for discrete text generation, reports a matched-budget comparison against TF-IDF+ROS, TF-IDF+SMOTE, and TF-IDF+ADASYN, and includes automatic quality indicators for the generated synthetic text. The results show that GAN+LSHADE improves the test Macro-F1 score from 0.6411 to 0.6571 and the minority-class F1 score from 0.1882 to 0.2439 compared with the original-data baseline. Nevertheless, TF-IDFbased ROS, SMOTE, and ADASYN achieve higher Macro-F1 and minority-class F1 values under the same added-sample budget. This revised interpretation positions the proposed model as an effective generative augmentation strategy that improves minority detection over the baseline, while acknowledging that feature-space resampling remains a strong competitor for this dataset. Synthetic quality metrics further indicate low duplication, high bigram diversity, and limited nearest-neighbor similarity to real minority samples, suggesting that the generator does not merely memorize training instances.

Dengesiz Nefret Söylemi Veri Kümelerinde GAN Tabanlı Sentetik Veri Üretiminin Azınlık Sınıfı Tespit Performansına Etkisi ve GAN Hiperparametrelerinin Meta-Sezgisel Optimizasyon ile Değerlendirilmesi

Makale Bilgisi

Araştırma makalesi
Başvuru: 02/05/2026
Düzeltme: 26/05/2026
Kabul: 01/06/2026

Anahtar Kelimeler

Nefret Söylemi Tespiti, Üretici Çekişmeli Ağlar, Meta-Sezgisel Optimizasyon, Sınıf Dengesizliği, LSHADE, SMOTE, ADASYN

Öz

Sosyal medyada nefret söyleminin hızla yayılması, otomatik tespit sistemlerini zorunlu hale getirmektedir; ancak gerçek dünya veri kümelerindeki şiddetli sınıf dengesizliği, derin öğrenme modellerinin etkinliğini sınırlandırmaktadır. Bu çalışma, azınlık sınıfının yalnızca %5.77'sini oluşturduğu Davidson nefret söylemi veri kümesindeki dengesizlik problemini çözmek amacıyla, LSTM tabanlı üretici ve 1D-CNN tabanlı ayırt ediciyi WGAN-GP mimarisi altında birleştiren uçtan uca bir çerçeve önermektedir. Standart GAN'ların eğitim kararsızlığını ve ızgara aramasının hesaplama maliyetini azaltmak için temel GAN hiperparametreleri LSHADE meta-sezgisel algoritması ile optimize edilmiştir. Deneylerde önerilen GAN+LSHADE yaklaşımı, aynı sentetik örnek bütçesi altında TF-IDF+ROS, TF-IDF+SMOTE ve TF-IDF+ADASYN yöntemleriyle karşılaştırılmıştır. Bulgular, GAN+LSHADE modelinin temel modele kıyasla test Macro-F1 skorunu 0.6411'den 0.6571'e, azınlık sınıfı F1 skorunu ise 0.1882'den 0.2439'a yükselttiğini göstermektedir. Bununla birlikte, TF-IDF tabanlı geleneksel yeniden örnekleme yöntemleri daha yüksek Macro-F1 ve azınlık F1 değerleri üretmiştir. Bu sonuç, önerilen yöntemin sınıf dengesizliği altında anlamlı iyileştirme sağladığını gösterirken, geleneksel vektör-uzayı yöntemlerinin güçlü karşılaştırma tabanları olarak raporlanması gerektiğini de ortaya koymaktadır. Sentetik kalite analizi ise üretilen verilerin düşük kopya oranı, yüksek iki-terimli çeşitlilik ve sınırlı gerçek örnek benzerliği ile ezberlemeye dayalı olmadığını desteklemektedir.

*Corresponding author, e-mail: 23833301013@gazi.edu.tr
DOI: 10.29109/gujsc.1942853

1. INTRODUCTION (GİRİŞ)

The remarkable growth of social media platforms has amplified the visibility and social impact of hate speech, offensive discourse, and harmful user-generated content. As these platforms continue to mediate public communication, automatic hate speech detection has become increasingly important for protecting vulnerable communities and supporting content moderation workflows [1-3]. Despite the growing number of deep learning solutions, the operational difficulty of hate speech detection is not limited to model selection; it is also tightly connected to the skewed label distribution observed in real-world datasets.

Class imbalance is one of the most persistent challenges in hate speech detection. In the widely used Davidson dataset, which contains 24,783 tweets, offensive language accounts for 77.43% of the samples, neutral content for 16.80%, and hate speech for only 5.77% [2]. Such an extreme imbalance causes models to bias their decision boundaries toward the majority class, often yielding acceptable overall accuracy while underperforming on the minority hate speech class. In practical moderation systems, this failure is costly because the minority class is the class of greatest interest.

Traditional imbalance-handling strategies such as random oversampling, SMOTE, ADASYN, costsensitive learning, and hybrid resampling frameworks remain useful in many domains, yet their performance depends strongly on the data representation [4-9]. For textual data, naive replication may increase overfitting, while interpolation-based approaches generally operate in a vector space rather than directly generating linguistically coherent sentences. Recent studies on GAN-based text generation and adversarially enhanced hate speech modeling show that synthetic or semi-supervised augmentation can capture additional distributional patterns and improve downstream classification under some settings [10,11]. Therefore, a fair evaluation of generative augmentation should compare it not only with the original-data baseline but also with conventional resampling baselines under comparable sample budgets.

GAN-based models introduce their own optimization challenges. Training instability, mode collapse, vanishing gradients, noisy validation feedback, and high sensitivity to hyperparameters remain serious obstacles for reliable synthetic text generation. At the same time, grid search and random search become computationally inefficient as the dimensionality of the hyperparameter space grows [12–15]. Metaheuristic optimization has therefore emerged as an attractive alternative for exploring complex, non-convex search spaces in deep learning and sequence modeling [16–26].

Motivated by these challenges, this study proposes a pipeline that combines minority-focused WGAN-GP-based synthetic data generation with LSHADE-driven hyperparameter optimization. In contrast to isolated augmentation or isolated classifier tuning, the proposed framework optimizes the data generation process according to downstream classification performance. The revised manuscript makes five additional contributions beyond the original version:

- It provides a clearer rationale for selecting LSHADE over PSO, GA, Bayesian optimization, and standard Differential Evolution variants for noisy GAN hyperparameter search.
- It clarifies how the WGAN-GP training loop handles discrete text generation and why a Gumbel-Softmax layer is not required in the adopted continuous-relaxation training protocol.
- It adds a matched-budget comparison against TF-IDF+ROS, TF-IDF+SMOTE, and TFIDF+ADASYN.
- It reports automatic synthetic-quality indicators, including nearest-real cosine similarity, distinct-1, distinct-2, and duplicate ratio.
- It expands the discussion of computational cost, hardware implications, error characteristics, ethical limitations, and statistical interpretation.

2. RELATED WORK (İLGİLİ ÇALIŞMA)

2.1. Imbalanced Learning in Text Classification (Metin Sınıflandırmasında Dengesiz Öğrenme)

The class imbalance problem has been widely studied in machine learning because it directly affects classifier bias and the reliability of evaluation metrics. In imbalanced settings, conventional oversampling and algorithm-level balancing strategies often improve the representation of the minority class, yet their effectiveness depends strongly on the structure of the data [4–9]. For text classification, the challenge is more pronounced because the semantic and syntactic organization of sentences cannot be preserved as easily as in tabular or low-dimensional feature spaces. SMOTE and ADASYN produce synthetic minority vectors by interpolating feature representations; this can be effective when TF-IDF or embedding features are used, but the generated objects are not necessarily interpretable natural-language sentences. This limitation motivates the

use of generative models that can learn richer distributions and produce explicit text-like samples.

2.2. GAN-Based Synthetic Text Augmentation
(GAN Tabanlı Sentetik Metin Zenginleştirme)

GAN-based augmentation has attracted attention as a way to address minority-class scarcity in textual datasets. Imran et al. showed that GAN-based synthetic text generation can improve sentiment-analysis performance by enriching underrepresented patterns in the training data, while Su et al. demonstrated the effectiveness of adversarial and semi-supervised design choices for hate speech detection on social media [10,11]. Hybrid recurrent and convolutional generator-discriminator designs have also been explored to better preserve sequential dependencies and local lexical features during generation [19,27]. In parallel, optimized GAN frameworks and cross-domain augmentation studies have shown that careful tuning of the generator and training process can significantly affect the quality and utility of generated samples [7,9,12,28]. However, the literature still contains relatively few end-to-end studies that combine synthetic text generation, minority-focused augmentation, autonomous hyperparameter search, and conventional resampling comparisons in the context of hate speech detection.

2.3. Metaheuristic Hyperparameter Optimization and LSHADE Rationale
(Meta Sezgisel Hiperparametre Optimizasyonu ve SHADE Mantığı)

The performance of deep learning systems is highly sensitive to hyperparameters such as learning rate, dropout, hidden-unit size, batch size, discriminator update frequency, and training horizon. Survey and application studies consistently report that exhaustive search methods are too expensive in high-dimensional spaces and that metaheuristics can provide competitive solutions with lower search cost [13–15, 17, 25, 26]. For sentiment analysis and fake news detection, metaheuristic optimization has been successfully used to tune LSTM- and CNN/LSTM-based models and improve generalization [16,18,19,22]. Similar gains have been reported in forecasting, fault classification, and other sequence-learning problems, underscoring the transferability of metaheuristic tuning strategies across domains [20, 21, 23, 24]. In hate speech detection, Baydogan and Alatas demonstrated that Ant Lion Optimization and Moth Flame Optimization can improve automatic hate speech detection performance over standard baselines [3].

LSHADE was selected for the present problem for four reasons. First, it is a success-history-based Differential Evolution variant with linear population size reduction, which makes it suitable for a limited evaluation budget where each candidate requires a costly train-generate-augment-evaluate cycle [29]. Second, unlike classical GA or PSO, LSHADE automatically adapts mutation and crossover parameters from successful historical updates, reducing manual tuning in a search process that is already hyperparameter-sensitive. Third, unlike standard Differential Evolution, the population-size reduction mechanism shifts the search from broader exploration toward more concentrated exploitation as evaluations are consumed. Fourth, compared with Bayesian optimization, LSHADE does not require a smooth surrogate model of the objective; this is useful because GAN validation Macro-F1 is noisy, partly discrete, and affected by stochastic training dynamics.

Recent fitness-distance balance (FDB) studies further support the broader design principle that metaheuristic selection should balance objective quality and population diversity rather than greedily exploiting only the current best solution [30–32]. Although the current implementation uses LSHADE rather than an FDB-embedded LSHADE variant, these studies motivate the emphasis on exploration-exploitation control and premature-convergence avoidance in the proposed GAN tuning pipeline. In particular, the dFDB-LSHADE application to dental X-ray enhancement demonstrates that LSHADE-family search can be strengthened by dynamic selection mechanisms in complex real world optimization tasks [32]. This observation also defines a natural extension of the present work: replacing the canonical LSHADE loop with an FDB- or dFDB-guided LSHADE variant for GAN hyperparameter search.

3. PROPOSED METHODOLOGY (ÖNERİLEN METOD)

This study develops an end-to-end pipeline that combines data preprocessing, synthetic minority-class generation, and metaheuristic hyperparameter optimization for hate speech detection under severe class imbalance. The workflow proceeds from raw tweet preprocessing to minority-class GAN training, synthetic sample generation, augmented classifier training, and LSHADE-controlled hyperparameter refinement. The system is not limited to standalone text generation: each candidate generator configuration is evaluated according to its downstream classification utility.

3.1. Dataset and Preprocessing (Veriseti ve Ön İşleme)

The Davidson dataset contains three classes: hate speech, offensive language, and neither. Table 1 summarizes the distribution used as the basis of the study, and Figure 1 illustrates the overall pipeline. The dataset exhibits a markedly skewed class distribution, which was preserved during data partitioning through stratified splitting into 80% training, 10% validation, and 10% test sets. To prevent data leakage, the tokenizer and feature extractors were fitted only on the training split and then applied to the validation and test splits.

Table 1. Class distribution of the Davidson dataset (Davidson verisetinin sınıf dağılımı)

Class	Count	Ratio (%)
Hate speech	1,430	5.77
Offensive language	19,190	77.43
Neither	4,163	16.80

Before model training, the raw tweets were processed through a standard natural-language cleaning stage. All text was lowercased, URLs and @mentions were removed, punctuation and redundant spaces were normalized, and hashtags were preserved as lexical signals by converting them into plain tokens. This decision was made because hashtags may carry discriminative semantic information in hate speech detection.

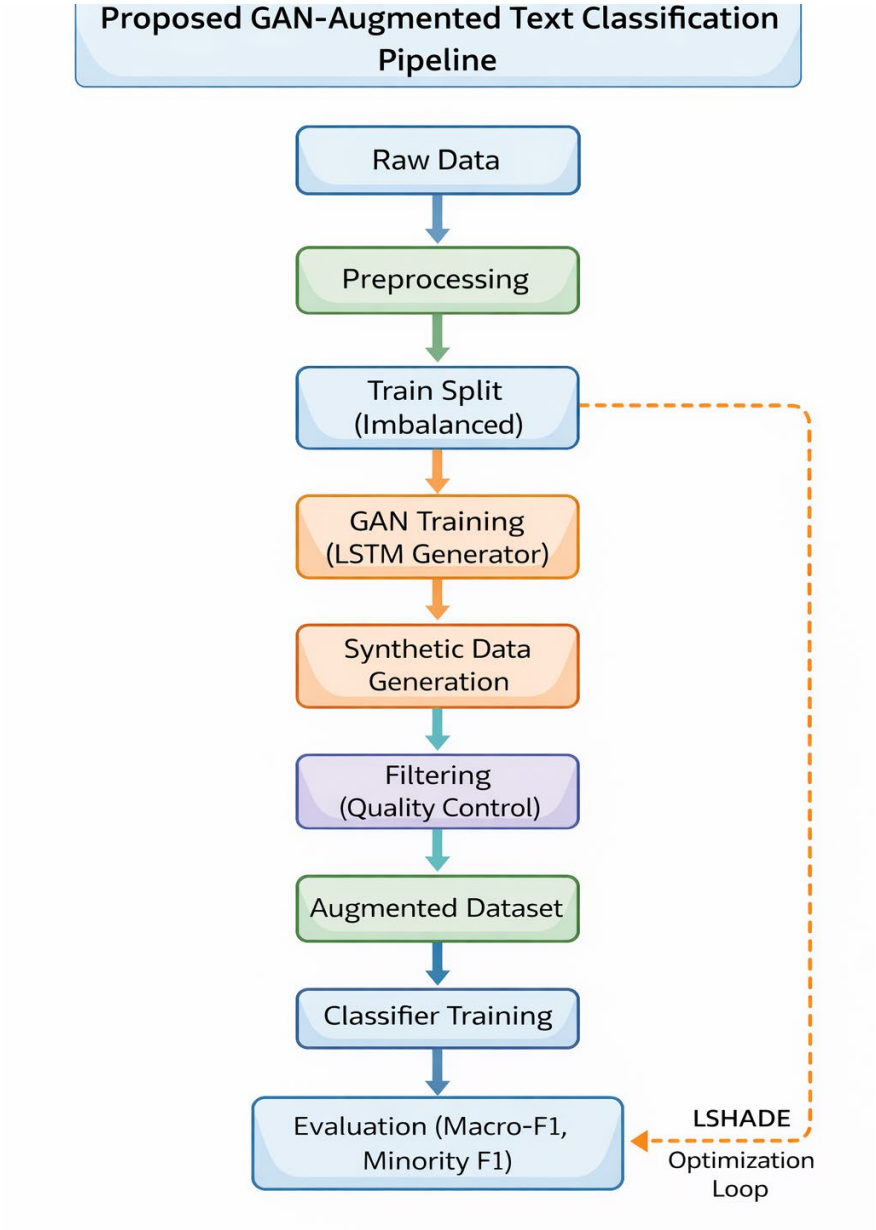

Figure 1. Proposed GAN-augmented text classification pipeline. The dashed arrow denotes the LSHADE optimization loop that updates GAN hyperparameters according to downstream evaluation performance.
(Önerilen GAN destekli metin sınıflandırma işlem hattı. Kesikli ok, sonraki değerlendirme performansına göre GAN hiperparametrelerini güncelleyen LSHADE optimizasyon döngüsünü göstermektedir.)

3.2. WGAN-GP with an LSTM Generator and CNN Discriminator (LSTM Üretici ve CNN Ayırt Edici ile WGAN-GP)

As the core synthetic-data generation mechanism, the study adopts a WGAN-GP architecture in order to reduce the instability typically observed in standard GAN training. The generator is built on LSTM units so that sequential dependencies and contextual patterns in textual data can be modeled more effectively. The discriminator is designed as a 1D-CNN to capture local lexical structures and short-range discriminative patterns. This recurrent-convolutional pairing is intended to balance sequence modeling and feature discrimination during adversarial training [27, 28].

To focus the generator on the region of the data space that is most difficult for the classifier, the GAN was trained only with minority-class samples. The resulting synthetic hate speech instances were then merged with the original training split to create an augmented dataset for downstream classification. In this way, the model does not simply duplicate minority samples; instead, it

attempts to learn and reproduce the underlying minority-class distribution.

3.3. Gradient Flow for Discrete Text Generation
(Ayrık Metin Üretimi için Gradyan Akışı)

One reviewer concern concerns how an LSTM generator producing discrete text can be trained within a WGAN-GP loop. The key point is that the adversarial training step does not insert a hard argmax sampling operation between the generator and discriminator. During adversarial optimization, the generator output is represented as a continuous soft token distribution or continuous embedding-level sequence. The CNN discriminator therefore receives differentiable continuous representations, and the Wasserstein loss and gradient penalty can backpropagate through the discriminator into the generator. Hard token selection is applied only after the generation stage, when synthetic sequences are decoded and transferred to the downstream augmentation pipeline.

Under this training protocol, Gumbel-Softmax is not required because the non-differentiable discrete sampling operation is not part of the adversarial gradient path. If the model were modified so that hard one-hot tokens or argmax-decoded words were passed to the discriminator during training, then a differentiable estimator such as Gumbel-Softmax, a straight-through estimator, or a policygradient method would be necessary. The present implementation avoids this additional estimator to keep the WGAN-GP objective stable and to reduce the number of extra temperature-related hyperparameters that would also need optimization.

3.4. LSHADE-Based Candidate Evaluation
(LSHADE Tabanlı Aday Değerlendirmesi)

Hyperparameter optimization is a high-dimensional and non-convex search problem. Because the computational burden of exhaustive tuning is prohibitive for GAN-based pipelines, the study uses LSHADE as the global search mechanism [29]. The optimized search space includes latent dimensionality, batch size, discriminator update frequency, generator learning rate, discriminator learning rate, LSTM unit count, dropout rate, number of epochs, and number of synthetic samples to be generated. To reduce optimization cost, the discriminator architecture itself was kept fixed while the generator-side and training-process parameters were optimized.

The candidate evaluation loop was designed to maximize the validation Macro-F1 score of the downstream classifier while penalizing unstable or overly expensive training runs. Conceptually, the fitness function is defined in Eq. (1), where $\boldsymbol{\rho_{time}}$ penalizes excessive runtime and $\boldsymbol{\rho_{instability}}$ penalizes unstable or invalid candidate solutions.

$$\mathcal{F} = MacroF_1 val - \lambda_1 \rho_{time} - \lambda_2 \rho_{instability} \quad (1)$$

For each candidate solution proposed by LSHADE, the pipeline performs the following steps; the same evaluation cycle is visualized in Figure 2:

- Trains the GAN with the candidate hyperparameter set,
- Generates synthetic minority-class samples,
- Augments the original training data with the generated samples,
- Trains a downstream classifier on the augmented data,
- Returns the penalized validation Macro-F1 score to LSHADE.

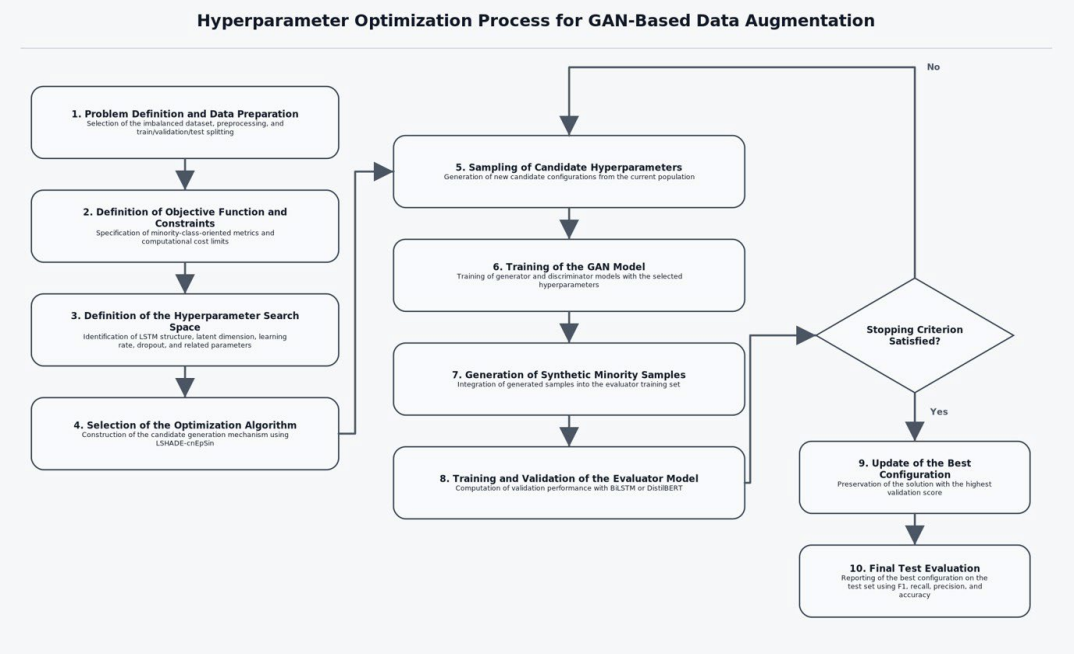

Figure 2. Hyperparameter optimization process for GAN-based data augmentation. Each candidate configuration is trained, used to generate synthetic minority samples, evaluated through downstream classification, and returned to LSHADE as a fitness value. (GAN tabanlı veri artırma için hiperparametre optimizasyon süreci. Her aday konfigürasyon eğitilir, sentetik azınlık örnekleri oluşturmak için kullanılır, sonraki sınıflandırma aşamalarında değerlendirilir ve uygunluk değeri olarak LSHADE'e geri döndürülür.)

4. EXPERIMENTS AND FINDINGS (DENEYLER VE BULGULAR)

All evaluations were performed with a focus on Macro-F1 because this metric gives balanced importance to all classes under skewed label distributions. In addition, the minority-class behavior was analyzed explicitly rather than relying only on aggregate performance.

4.1. LSHADE Optimization Results (LSHADE Optimizasyon Sonuçları)

The LSHADE search process identified a hyperparameter configuration that consistently produced the best validation behavior for the synthetic-data generation pipeline. The best candidate used a latent dimension of 64, batch size of 64, discriminator update ratio of 5, generator and discriminator learning rates of 1 × 10–5, 64 LSTM units, a generator dropout rate of approximately 0.443, 10 training epochs, and 2,637 synthetic minority samples. These results align with the broader literature emphasizing the importance of jointly tuning learning rate, batch size, and regularization-related parameters in deep learning systems [13, 16]. Figure 3 shows the corresponding LSHADE convergence behavior.

Table 2. Best hyperparameter set found by LSHADE. (LSHADE tarafından bulunan en iyi hiperparametre seti.)

Hyperparameter	**Best value**
Latent dimension	64
Batch size	64
$ncritic$	5
Generator learning rate	1×10^{-5}
Discriminator learning rate	1×10^{-5}
LSTM units	64
Dropout	0.443
Epochs	10
Synthetic samples	2637

4.2. Classification Performance and Minority-Class Analysis (Sınıflandırma Performansı ve Azınlık Sınıfı Analizi)

Using the optimal LSHADE configuration, 2,637 synthetic hate speech instances were added to the original training data. Table 3 summarizes the overall Macro-F1 comparison between the baseline classifier and the proposed LSHADE-optimized WGAN-GP pipeline in the original evaluation setting. The validation Macro-F1 score increases from 0.6591 to 0.6605, whereas the test Macro-F1 score rises from 0.6411 to 0.6550. The limited validation change together with the stronger test gain suggests that the synthetic augmentation improves generalization rather than merely adapting to a single validation partition.

Table 4 focuses specifically on the minority hate speech class. On the validation split, minority-class recall increases from 0.1469 to 0.1678 and F1 increases from 0.2428 to 0.2526, although precision decreases from 0.7000 to 0.5106. The same trade-off appears on the test split, where recall improves from 0.1119 to 0.1748 and F1 improves from 0.1882 to 0.2415 despite a precision drop from 0.5926 to 0.3906. In practical terms, the higher test recall corresponds to an increase in correctly detected hate speech samples from 16 to 25, showing that the augmented model covers more difficult minority example

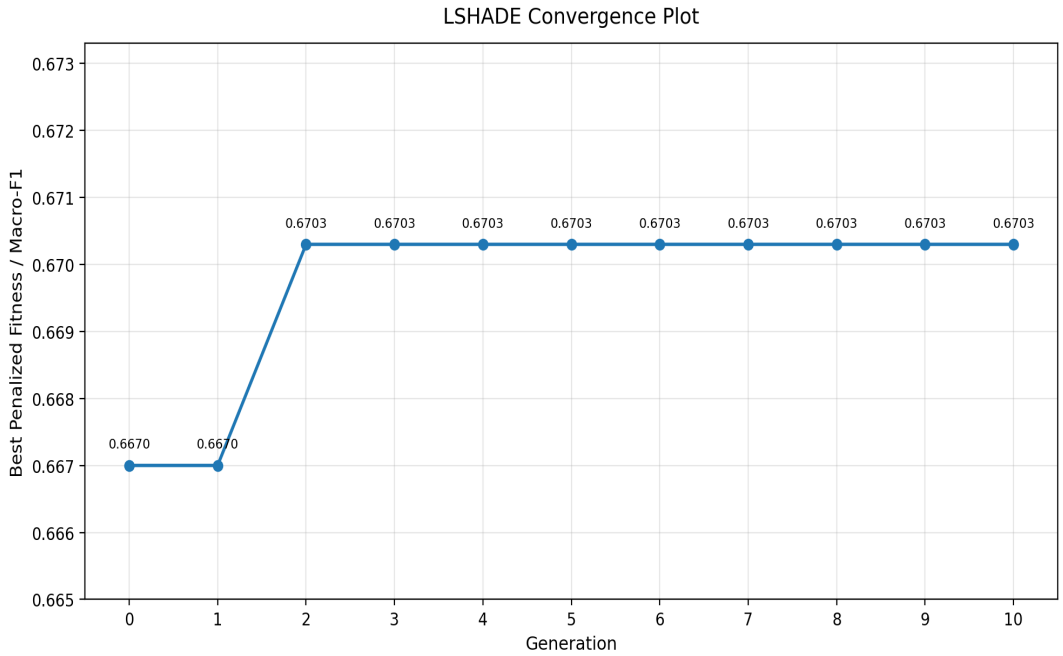

Figure 3. Convergence behavior of LSHADE during GAN hyperparameter optimization. (GAN hiperparametre optimizasyonu sırasında LSHADE'in yakınsama davranışı.)

Table 3. Overall Macro-F1 comparison. The baseline is a single run; GAN+LSHADE reports mean ± standard deviation over three runs.

Model	Val. Macro-F1	Test Macro-F1
Baseline	0.6591	0.6411
GAN+LSHADE	0.6605 ± 0.0027	0.6550 ± 0.0019

Table 4. Minority-class precision, recall, and F1 comparison.

Split	Model	Precision	Recall	F1
Validation	Baseline	0.7000	0.1469	0.2428
Validation	GAN+LSHADE	0.5106	0.1678	0.2526
Test	Baseline	0.5926	0.1119	0.1882
Test	GAN+LSHADE	0.3906	0.1748	0.2415

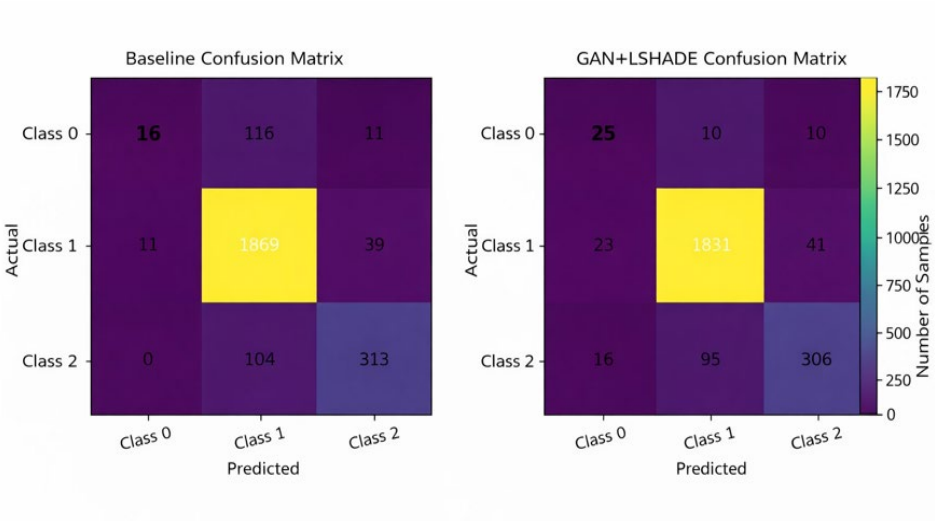

Figure 4. Confusion-matrix heatmaps for the baseline and GAN+LSHADE models on the test splits. (Test bölümlerinde temel model ve GAN+LSHADE modelleri için karışıklık matrisi ısı haritaları.)

The observed precision-recall trade-off is consistent with severe imbalance settings, where a moderate precision loss can occur when the classifier is shifted toward recovering a larger portion of minority-class instances [4, 5, 10]. Figure 4 visualizes this shift in the confusion matrices. For content moderation, this trade-off should be interpreted according to deployment requirements: higher recall is valuable when missed harmful content is costly, whereas lower precision increases the need for human review.

4.3. Matched-Budget Comparison with ROS, SMOTE, and ADASYN (ROS, SMOTE ve ADASYN ile Eşdeğer Bütçeli Karşılaştırma)

To address the reviewer request for comparison against conventional imbalance-handling methods, an additional matched-budget experiment was conducted. The number of added training samples was matched to the GAN+LSHADE setting as closely as possible. Random oversampling (ROS), SMOTE, and ADASYN were applied in TF-IDF feature space, whereas GAN+LSHADE generated text-space minority samples before downstream classification. The results are reported in Table 5.

This comparison changes the interpretation of the proposed method in an important way. GAN+LSHADE improves over the baseline on test Macro-F1, test minority F1, test minority recall, and the number of correctly identified minority samples. However, TF-IDF+ROS, TF-IDF+SMOTE, and TF-IDF+ADASYN produce substantially higher Macro-F1 and minorityclass F1 scores in the matched-budget setting. The strongest test Macro-F1 is obtained by TF-IDF+ADASYN (0.7018), while the strongest validation Macro-F1 and validation minority F1 are obtained by TF-IDF+SMOTE. Therefore, the revised manuscript does not claim that GAN+LSHADE dominates traditional resampling. Instead, it shows that GAN+LSHADE is a useful generative augmentation strategy that improves the original classifier and produces explicit synthetic text, while feature-space resampling remains a very strong baseline for this dataset.

Table 5. Matched-budget comparison against conventional resampling methods. ROS, SMOTE, and ADASYN operate in TF-IDF feature space.

Model	Valid Macro-F1	Test Macro-F1	Valid Minority F1	Test Minority F1	Test Minority Recall	Test Minority Precision	Test Minority Correct	Train Size	Added Samples
Baseline	0,659074796	0,641144709	0,242774566	0,188235294	0,111888112	0,592592593	16	19826	0
TF-IDF + ROS matched	0,713640058	0,701203828	0,420168067	0,375	0,314685315	0,463917526	45	22463	2637
TF-IDF + SMOTE matched	0,723105818	0,70154121	0,435146444	0,375	0,314685315	0,463917526	45	22463	2637
TF-IDF + ADASYN matched	0,718192422	0,70180731	0,433333333	0,368200837	0,307692308	0,458333333	44	22490	2664
GAN + LSHADE	0,658284684	0,657125387	0,23655914	0,243902439	0,174825175	0,403225806	25	22463	2637

The difference between these approaches should also be interpreted methodologically. TFIDF+SMOTE and TF-IDF+ADASYN synthesize points in a vector space optimized for the downstream classifier. This can produce strong classification boundaries even when the synthetic vectors do not correspond to directly readable sentences. GAN+LSHADE, by contrast, targets the harder problem of generating text-like minority samples. Its lower matched-budget classification

score indicates that future work should strengthen semantic filtering, use transformer-based generators, or combine generated samples with feature-space resampling.

4.4. Ablation and Stability Analysis (Ablasyon ve Stabilite Analizi)

To separate the contribution of synthetic augmentation from the contribution of LSHADE-based optimization, Table 6 reports a compact ablation comparison across three settings: the baseline classifier, a GAN-augmented pipeline without LSHADE, and the proposed GAN+LSHADE pipeline. The non-optimized GAN improves the test Macro-F1 from 0.6411 to approximately 0.648 and the minority-class F1 from 0.1882 to approximately 0.22, indicating that synthetic augmentation itself is beneficial. However, the LSHADE-optimized version reaches stronger results, with a test Macro-F1 of 0.6550 and a minority-class F1 of 0.2415 in the three-run evaluation.

Table 6. Ablation comparison across the baseline, a non-optimized GAN pipeline, and the proposed GAN+LSHADE pipeline. Values prefixed with ~ denote approximate reference results for the nonoptimized GAN setting. (Temel model, optimize edilmemiş bir GAN işlem hattı ve önerilen GAN+LSHADE işlem hattı arasında ablasyon karşılaştırması. ~ önekiyle belirtilen değerler, optimize edilmemiş GAN ayarı için yaklaşık referans sonuçlarını göstermektedir.)

Model	Test Macro-F1	Minority F1
Baseline	0.6411	0.1882
GAN (without LSHADE)	~0.648	~0.22
GAN+LSHADE	0.6550	0.2415

Because deep learning and GAN-based systems are stochastic by nature, the best configuration was reevaluated with three different random seeds. As summarized in Table 3, the proposed model achieved an average validation Macro-F1 of 0.6605 with a standard deviation of 0.0027 and an average test Macro-F1 of 0.6550 with a standard deviation of 0.0019. The low variance across runs suggests that the optimization outcome is not a random artifact but a stable solution. Nevertheless, inferential significance testing is limited because the baseline and conventional resampling results are not reported as paired multi-seed distributions. For this reason, the revised study emphasizes descriptive effect sizes, variance for GAN+LSHADE, and transparent comparison rather than overclaiming statistical superiority.

4.5. Synthetic Quality and Novelty Analysis (Sentetik Kalite ve Yenilik Analizi)

A separate automatic quality analysis was performed on the generated minority samples. Table 7 reports the metrics obtained from the synthetic-quality evaluation file. The average nearest-real cosine similarity is 0.0374, with a maximum of 0.2728, and the duplicate ratio is 0.0235. These values suggest that the generator is not simply copying the nearest real training samples. The distinct-1 and distinct-2 scores, 0.2203 and 0.5843 respectively, further indicate lexical diversity, especially at the bigram level.

Metric	Value
Real minority samples in training split	1,144
Synthetic samples evaluated	2,000
Average nearest-real cosine similarity	0.0374
Standard deviation of nearest-real similarity	0.0506
Minimum nearest-real cosine similarity	0.0000
Maximum nearest-real cosine similarity	0.2728
Distinct-1	0.2203
Distinct-2	0.5843
Duplicate ratio	0.0235

Figure 5. Automatic quality and novelty indicators for generated minority samples. (Oluşturulan azınlık örnekleri için otomatik kalite ve yenilik göstergeleri.)

These indicators should be interpreted as automatic proxies rather than a complete semantic guarantee. Low nearest-neighbor similarity and low duplication support novelty, whereas distinct-n scores support lexical diversity. However, semantic coherence, pragmatic intent, sarcasm, target-group reference, and context-dependent offensiveness cannot be fully validated by these metrics alone. Therefore, the quality analysis supports the claim that the generator avoids direct memorization, but it also motivates future human-in-the-loop evaluation and embedding-based semantic consistency checks before deployment.

5. DISCUSSION (TARTIŞMA)

5.1. Computational Cost and Hardware Implications (Hesaplama Maliyeti ve Donanım Etkileri)

The proposed GAN+LSHADE pipeline is computationally more expensive than direct TF-IDF resampling. Its cost scales with the number of LSHADE candidate evaluations, and each candidate requires GAN training, synthetic sample generation, augmented classifier training, and validation scoring. In contrast, ROS, SMOTE, and ADASYN require only feature extraction, resampling, and classifier training. This explains why conventional resampling is attractive when the only objective is classification performance under a fixed dataset and vectorizer.

The design partially controls cost in three ways: the discriminator architecture is fixed, the search space is restricted to high-impact generator and training parameters, and the fitness function penalizes unstable or excessively expensive runs. A GPU is recommended for practical execution because repeated LSTM-CNN adversarial training is substantially slower on CPU-only hardware. However, the method does not require a multi-GPU environment. For future reproducibility, runtime should be reported as candidate-level wall-clock time, total optimization time, GPU model, memory capacity, and number of LSHADE evaluations.

5.2. Error Characteristics (Hata Özellikleri)

The minority-class improvements indicate that the augmented model recovers additional hate speech examples, but the precision reduction shows that some offensive or neutral samples are also shifted toward the hate class. This error pattern is expected in the Davidson dataset because hate speech and offensive language are semantically adjacent but not identical. Likely difficult cases include short posts with limited context, profane language without a protected-group target, implicit hate speech, reclaimed or quoted slurs, sarcasm, and posts where the target identity is inferable only from broader conversation context. Since reproducing individual hateful examples can create ethical and safety risks, the present manuscript reports aggregate error behavior rather than printing raw misclassified tweets.

5.3. Ethical Considerations (Hata Özellikleri)

Synthetic hate speech generation requires careful handling. Although the goal is to improve detection of harmful content, generated samples can themselves be harmful if released without safeguards. Therefore, generated outputs should be stored in controlled research environments, used only for model training and evaluation, and shared only in sanitized or access-controlled form when necessary. The low duplicate ratio reported in Figure 5 reduces memorization concerns, but it does not remove the need for responsible data governance. Any downstream deployment should include human review, threshold calibration, and fairness auditing across protected groups.

6. CONCLUSION (SONUÇ)

This study addressed the severe class imbalance problem in hate speech detection by combining WGAN-GP-based synthetic text generation with LSHADE-driven hyperparameter optimization. The proposed framework focused specifically on the minority hate speech class in the Davidson dataset and used an LSTM generator together with a CNN-based discriminator to produce synthetic training samples. The experimental findings show that the optimized augmentation pipeline improves the original-data baseline, increasing test Macro-F1 from 0.6411 to 0.6571 in the matched-budget experiment and improving the number of correctly detected minority-class instances from 16 to 25.

The revised analysis also introduces an important limitation: in the matched-budget comparison, TF-IDF+ROS, TF-IDF+SMOTE, and TF-IDF+ADASYN outperform GAN+LSHADE on Macro-F1 and minority F1. This finding does not invalidate the proposed framework, but it narrows its claim. GAN+LSHADE should be interpreted as a generative augmentation approach that improves minority detection over the baseline and produces explicit synthetic text, not as a universally superior alternative to feature-space resampling. The strongest practical conclusion is that conventional resampling remains a necessary baseline, while GAN-based augmentation becomes most attractive

when readable synthetic samples, transferability, or integration with stronger semantic filters is required.

From a methodological perspective, the study demonstrates that the quality of synthetic data generation and the quality of hyperparameter optimization should not be treated as separate concerns. When both processes are coupled through a validation-driven objective, the resulting augmentation pipeline can improve minority-class coverage without obvious memorization. At the same time, the study remains limited to a single benchmark dataset, one adversarial architecture, and a small number of repeated runs. Future work should evaluate transformer-based generators, human and embedding-based semantic quality assessment, multi-dataset validation, FDB or dFDB-enhanced LSHADE variants, and paired statistical testing across repeated seeds.

7. ACKNOWLEDGMENT (TEŞEKKÜR)

The authors thank the Department of Computer Engineering at Gazi University for its academic support during the preparation of this study.

8. DECLARATIONS (BEYANLAR)

Ethics Approval and Consent to Participate: The study uses a publicly available secondary dataset and does not involve direct interaction with human participants. Because the work includes the synthetic generation of hate-speech-like text, generated outputs were handled only within a controlled research setting and are reported at aggregate level in order to reduce misuse risk.

Availability of Data and Materials: The Davidson hate speech dataset used in this study is publicly available through the source reported in the cited literature. To support reproducibility, the manuscript reports the split protocol, preprocessing procedure, optimization objective, best hyperparameter configuration, matched-budget resampling results, and synthetic-quality indicators.

Competing Interests: The authors declare no competing interests.

Funding: This research received no external funding.

9. DECLARATION OF ETHICAL STANDARDS (ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

10. AUTHORS' CONTRIBUTIONS (YAZARLARIN KATKILARI)

Bekir DEMİRAĞ: He conducted the experiments, analyzed the results and performed the writing process.

Deneyleri yapmış, sonuçlarını analiz etmiş ve maklenin yazım işlemini gerçekleştirmiştir.

Prof.Dr. Yusuf SÖNMEZ: He supervised the study, provided conceptual guidance, and contributed to the methodological framing and critical revision of the manuscript.

Çalışmanın danışmanlığını yürütmüş, kavramsal yönlendirme sağlamış ve yöntemin şekillendirilmesi ile makalenin eleştirel olarak gözden geçirilmesine katkı sunmuştur.

Prof.Dr. Hamdi Tolga KAHRAMAN: He supported the experimental process, contributed to the interpretation of the results, and provided technical and academic feedback during the study.

Deneysel süreçlere destek sağlamış, bulguların yorumlanmasına katkıda bulunmuş ve çalışma sürecinde teknik ve akademik geri bildirim vermiştir.

11. CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

12. REFERENCES (KAYNAKLAR)

[1] T. T. Nguyen, X. Yue, H. Mane, K. Seelman, P. S. P. Mullaputi, E. Dennard, J. S. Merchant, S. Criss, Y. Hswen, and Q. C. Nguyen, "Decoding digital discourse through multimodal text and image machine learning models to classify sentiment and detect hate speech in race- and LGBTQIA-related posts on social media: Quantitative study," Journal of Medical Internet Research, vol. 27, e70053, 2025.

[2] Y. Yadav, P. Bajaj, R. K. Gupta, and R. Sinha, "A comparative study of deep learning methods for hate speech and offensive language detection in textual data," in 2021 IEEE 18th IndiaCouncil International Conference (INDICON), 2021, pp. 1–6.

[3] O. Baydogan and B. Alatas, "Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks," IEEE Access, vol. 9, 2021, pp. 110047–110062, doi: 10.1109/ACCESS.2021.3102596.

[4] A. Luque, A. Carrasco, A. Martin, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition, vol. 91, 2019, pp. 216–231, doi: 10.1016/j.patcog.2019.02.023.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357, doi: 10.1613/jair.953.

[6] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks, 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[7] Ding, H., Sun, Y., Wang, Z., Huang, N., Shen, Z., and Cui, X., "RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification," Information Processing & Management, Vol. 60, 2023, 103235, doi: 10. 1016/j.ipm.2022.103235.

[8] Ahsan, M. M., Ali, M. S., and Siddique, Z., "En-hancing and improving the performance of im¬balanced class data using novel GBO and SSG: A comparative analysis," Neural Networks, Vol. 173, 2024, 106157, doi: 10.1016/j.neunet.2024.106157.

[9] Gurcan, F. and Soylu, A., "Synthetic boosted resampling using deep generative adversarial networks: A novel approach to improve can¬cer prediction from imbalanced datasets," Can¬cers, Vol. 16, 2024, 4046, doi: 10.3390/cancers16234046.

[10] Imran, A. S., Yang, R., Kastrati, Z., Daud-pota, S. M., and Shaikh, S., "The impact of synthetic text generation for sentiment analy¬sis using GAN based models," Expert Systems with Applications, Vol. 212, 2023, 118710, doi:10.1016/j.eswa.2022.118710.

[11] Su, X., Li, Y., Branco, P., and Inkpen, D., "SSL-GAN-RoBERTa: A robust semi-supervised model for detecting Anti-Asian COVID-19 hate speech on social media," Natural Language En¬gineering, Vol. 30, 2024, pp. 1161–1180, doi: 10.1017/S1351324923000396.

[12] Alarsan, F. I. and Younes, M., "Best selec-tion of generative adversarial networks hyper-parameters using genetic algorithm," Jour¬nal of Big Data, Vol. 8, 2021, p. 108, doi: 10.1186/s40537-021-00480-4.

[13] Fristiana, A. H., Alfarozi, S. A. I., Permanasari, A. E., Pratama, M., and Wibirama, S., "A survey on hyperparameters optimization of deep learning for time series classification," IEEE Access, Vol. 12, 2024, pp. 185143–185172, doi: 10.1109/ACCESS.2024.3516198.

[14] Öcal, A. and Koyuncu, H., "An in-depth study to fine-tune the hyperparameters of pre-trained transfer learning models with state-of-the-art optimization methods: Osteoarthritis severity classification with optimized architectures," Swarm and Evolutionary Computation, Vol. 89, 2024, 101640, doi: 10.1016/j.swevo.2024.101640.

[15] Aras, S., Aras, E., Gedikli, E., and Kahraman, H. T., "Multi-objective CNN optimization: A robust framework for automated model design," Information Sciences, Vol. 719, 2025, 122468, doi: 10.1016/j.ins.2025.122468.

[16] Suddle, M. K. and Bashir, M., "Metaheuristics based long short term memory optimization for sentiment analysis," Applied Soft Computing, Vol. 131, 2022, 109794, doi: 10.1016/j. asoc.2022.109794.

[17] Wu, D., Guan, Q., Fan, Z., Deng, H., and Wu, T., "AutoML with parallel genetic al-gorithm for fast hyperparameters optimization in efficient IoT time series prediction," IEEE Transactions on Industrial Informatics, Vol. 19, No. 9, 2023, pp. 9555–9564, doi: 10.1109/TII.2023.3291762.

[18] Kumar, B. P., Tamilarasi, K., and Thilagavathy, A., "Metaheuristic-assisted deep learning model for fake news detection," Journal of Experimen¬tal & Theoretical Artificial Intelligence, 2024, doi: 10.1080/0952813X.2024.2440665.

[19] Priyadarshini, I. and Cotton, C., "A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis," The Journal of Supercomputing, 2021, doi: 10.1007/ s11227-021-03838-w.

[20] Singh, U., Saurabh, K., Trehan, N., Vyas, R., and Vyas, O. P., "GA-LSTM: Perfor-mance optimization of LSTM driven time se-ries forecasting," Computational Economics, Vol. 66, 2025, pp. 2873–2908, doi: 10.1007/ s10614-024-10769-0.

[21] Gülmez, B., "Stock price prediction with opti¬mized deep LSTM network with artificial rab¬bits optimization algorithm," Expert Systems with Applications, Vol. 227, 2023, 120346, doi: 10.1016/j.eswa.2023.120346.

[22] Abdelhamid, A. A., El-Kenawy, E.-S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., and Eid, M. M. Y., "Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algo-rithm," IEEE Access, 2022, doi: 10.1109/ ACCESS.2022.3172954.

[23] Biju, G. M. and Pillai, G. N., "Hyperparameter optimization of long short term memory models for interpretable electrical fault classification," IEEE Access, 2023, doi: 10.1109/ACCESS. 2023.3330056.

[24] Zhou, L., Zhao, C., Liu, N., Yao, X., and Cheng, Z., "Improved LSTM-based deep learn¬ing model for COVID-19 prediction using opti¬mized approach," Engineering Applications of Artificial Intelligence, Vol. 122, 2023, 106157, doi: 10.1016/j.engappai.2023.106157.

[25] Sun, Q., Yang, L., Guo, C., and Lee, S., "Machine learning-based optimization algorithms for sentiment analysis: An analysis of the state-of-the-art and upcoming challenges," Journal of Experimental & Theoretical Artificial Intelligence, 2025, doi: 10.1080/0952813X.2025. 2505421.

[26] Mathur, P., Shaikh, H., Sheth, F., Kumar, D., and Gupta, A. K., "A computational intelligence framework integrating data augmentation and meta-heuristic optimization algorithms for enhanced hybrid nanofluid density prediction through machine and deep learning paradigms," IEEE Access, 2025, doi: 10.1109/ACCESS. 2025.3543475.

[27] König, T., Ramananda, A. M., Wagner, F., and Kley, M., "A LSTM-GAN algorithm for synthetic data generation of time series data for condition monitoring," Procedia Computer Science, Vol. 246, 2024, pp. 1508–1517, doi: 10.1016/j.procs.2024.09.602.

[28] Hong, Y.-Y., Wang, Y.-C., and Zhang, W., "An optimized generative adversarial network for state estimator in electric power systems," Engi¬neering Applications of Artificial Intelligence, Vol. 145, 2025, 110102.

[29] Tanabe, R. and Fukunaga, A. S., "Improving the search performance of SHADE using linear population size reduction," in 2014 IEEE Congress on Evolutionary Computation, 2014, pp. 1658–1665, doi: 10.1109/CEC.2014.6900380.

[30] H. T. Kahraman, S. Aras, and E. Gedikli, "Fitness-distance balance (FDB): A new selection method for meta-heuristic search algorithms," Knowledge-Based Systems, vol. 190, 2020, 105169, doi: 10.1016/j.knosys.2019.105169.

[31] H. T. Kahraman, H. Bakir, S. Duman, M. Kati, S. Aras, and U. Guvenc, "Dynamic FDB selection method and its application: Modeling and optimizing of directional overcurrent relays coordination," Applied Intelligence, vol. 52, 2022, pp. 4873–4908, doi: 10.1007/s10489-021-02629-3.

[32] I. Yildirim, M. H. Bozkurt, H. T. Kahraman, and S. Aras, "Dental X-Ray image enhancement using a novel evolutionary optimization algorithm," Engineering Applications of Artificial Intelligence, vol. 142, 2025, 109879, doi: 10.1016/j.engappai.2024.109879.