Towards Scalable Monitoring: A Robust Few-Shot Multimodal Framework for Migration Detection on TikTok

Dimitrios Taranis *, Gerasimos Razis, Ioannis Anagnostopoulos

Posted Date: 19 January 2026

doi: 10.20944/preprints202601.1402.v1

Keywords: TikTok; multimodal classification; irregular migration; social media analysis; video understanding; OCR; content moderation; weak supervision; interpretable feature fusion

 doi:10.20944/preprints202601.1402.v1

Article

Towards Scalable Monitoring: A Robust Few-Shot Multimodal Framework for Migration Detection on TikTok

Dimitrios Taranis *, Gerasimos Razis and Ioannis Anagnostopoulos

Department of Computer Science and Biomedical Informatics, University of Thessaly, 35131 Lamia, Greece

* Correspondence: dtaranis@uth.gr

Abstract

Short-form video platforms such as TikTok host large volumes of user-generated, often ephemeral, content related to irregular migration, where relevant cues are distributed across visual scenes, on-screen text, and multilingual captions. Automatically identifying migration-related videos is challenging due to this multimodal complexity and the scarcity of labeled data in sensitive domains. This paper presents an interpretable few-shot multimodal classification framework designed for deployment under data-scarce conditions. We extract features from platform metadata, automated video analysis (Google Cloud Video Intelligence), and Optical Character Recognition (OCR) text, and compare text-only, OCR-only, and vision-only baselines against a multimodal fusion approach using Logistic Regression, Random Forest, and XGBoost. In this pilot study, multimodal fusion consistently improves class separation over single-modality models, achieving an F1-score of 0.92 for the migration-related class under stratified cross-validation. Given the limited sample size, these results are interpreted as evidence of feature separability rather than definitive generalization. Feature importance and SHAP analyses identify OCR-derived keywords, maritime cues, and regional indicators as the most influential predictors. To assess robustness under data scarcity, we apply SMOTE to synthetically expand the training set to 500 samples and evaluate performance on a small held-out set of real videos, observing stable results that further support feature-level robustness. Finally, we demonstrate scalability by constructing a weakly labeled corpus of 600 videos using the identified multimodal cues, highlighting the suitability of the proposed feature set for weakly supervised monitoring at scale. Overall, this work serves as a methodological blueprint for building interpretable multimodal monitoring pipelines in sensitive, low-resource settings.

Keywords: TikTok; multimodal classification; irregular migration; social media analysis; video understanding; OCR; content moderation; weak supervision; interpretable feature fusion

1. Introduction

Short-form video platforms such as TikTok have become central arenas for sharing content related to migration journeys, border crossings, and humanitarian crises[1,2]. Videos in this context are typically multimodal, combining visual scenes, on-screen text, audio, and metadata in ways that complicate automated analysis. Identifying which videos depict or reference irregular migration is particularly challenging when critical cues appear in overlaid text or visual patterns, while captions are short, multilingual, or noisy. Existing computational research on social media often focuses on text-only content or static images, leaving the joint modeling of video signals, OCR text, and contextual metadata comparatively underexplored for migration-related narratives[3,4]. Rather than pursuing end-to-end deep architectures, this study deliberately adopts a feature-based, interpretable modeling strategy. This choice reflects the constraints of sensitive domains such as irregular migration, where data availability is limited, annotation is costly, and transparency of automated decisions is essential. The goal is therefore not to maximize raw predictive performance, but to

 doi:10.20944/preprints202601.1402.v1

understand which multimodal cues consistently signal migration-related narratives under realistic data-scarcity conditions. Rather than competing with end-to-end multimodal architectures, this work addresses a complementary question: which interpretable multimodal cues remain reliable under extreme data scarcity.

1.1. Background

Prior work on multimodal classification of social media posts has shown that combining textual and visual features can substantially improve the detection of events, misinformation, and crisis-related content compared to single-modality approaches[4,5]. Studies of TikTok in particular highlight the platform's complexity, with fast-paced editing, on-screen captions, and diverse creator communities that make conventional text-based pipelines insufficient. For sensitive domains such as migration, this multimodal complexity intersects with ethical and practical constraints on data access, annotation, and sharing, which in turn limits the availability of public, well-documented datasets[6,7].

1.2. Motivation

In the context of irregular migration, TikTok videos may depict boats, sea crossings, group movement, or coastal regions that implicitly signal migration routes, while captions are often in Arabic or other non-English languages[7]. As a result, moderation and research workflows that rely only on captions or hashtags risk missing relevant content or misclassifying ambiguous posts. At the same time, labeled datasets in this space remain scarce and typically small-scale, making it difficult to benchmark multimodal approaches and to understand which signals are most informative for migration-related classification. There is therefore a need for pilot studies that systematically compare modalities and extract interpretable insights that can inform larger-scale, weakly supervised labeling in future work[8–10].

Building on our prior research on irregular migration discourse on Twitter[11] and its evolution across X (formerly Twitter) and Telegram[12], this study addresses the significantly higher complexity of the TikTok video landscape. While our previous work demonstrated the efficacy of text-based Transformer models in low-resource settings, short-form video presents a dual challenge: relevant cues are dispersed across modalities (visual, audio, OCR), and the content itself is highly ephemeral due to platform moderation.

1.3. Contributions

This paper offers a small-scale, multimodal study of migration-related content detection on TikTok using interpretable feature fusion. First, it introduces a manually annotated dataset of 50 videos (35 migration-related, 15 non-related) enriched with features derived from platform metadata, automated video analysis, and OCR text extracted from video frames and images. Second, it benchmarks text-only, OCR-only, and vision-only baselines against a fusion configuration using Logistic Regression, Random Forest, and XGBoost classifiers, showing that multimodal fusion consistently outperforms single-modality models in stratified cross-validation. Third, it provides feature-importance and SHAP analyses that highlight the central role of OCR-derived keywords, maritime labels, and regional indicators for migration-related classification, and discusses how these findings can guide future weakly supervised labeling at larger scale.

In summary, the main contributions of this work are:

- A lightweight, interpretable multimodal fusion framework: We propose an interpretable classification pipeline that combines text, OCR, and visual signals to detect migration-related content on TikTok, outperforming single-modality baselines.
- Identification of robust predictors: Through SHAP and feature importance analysis, we demonstrate that OCR-derived keywords (e.g., route terms) and regional indicators are the most reliable signals for detection, offering higher discriminative power than generic visual labels.

 doi:10.20944/preprints202601.1402.v1

- Robustness verification via synthetic augmentation: We validate the stability of the proposed features by scaling the training set to 500 samples using the Synthetic Minority Over-sampling Technique (SMOTE), providing evidence that the high classification performance is robust and not merely an artifact of the small pilot dataset.

1.4. Related Work

Research on migration and social media has shown how platforms such as TikTok, Instagram, and Facebook enable migrants and their families to share experiences, negotiate belonging, and exchange information about routes and risks, while also being used by smugglers to advertise crossings and connect with potential clients[13,14]. Studies of hashtags and communities (e.g. work on #Migrantes) highlight how platform-specific affordances and recommendation systems shape the visibility of precarious migration narratives, including emerging forms of "digital smuggling" and the circulation of route information via short videos and comments. Recent studies on TikTok 'refugees' communities[15] highlight detection challenges our features address[16]. In parallel, a growing body of work examines TikTok's multimodal complexity, proposing annotation schemes and analytical frameworks that take into account the joint role of video, audio, and on-screen text in meaning-making. Recent Transformer-based systems such as MTikGuard detect harmful content on TikTok with high performance (reported F1 ≈ 0.89), but are designed for large-scale moderation and do not focus on interpretable migration-specific cues[17].

Beyond migration, multimodal classification approaches have reported consistent performance gains by fusing textual and visual representations. For example, Sánchez-Villegas et al.[18] demonstrate improved multimodal classification performance using Ber-ViT fusion on large-scale social media datasets. In contrast, our work focuses on interpretable feature-level fusion under extreme data scarcity. These studies frequently rely on deep architectures and large datasets, but they rarely focus on small, highly sensitive domains where interpretability and careful feature design are essential. Beyond academic work, industry and technical reports emphasize that multimodal moderation can uncover harmful or policy-violating content that text-only systems miss, by jointly analyzing captions, audio, and visual scenes[19–21]. This broader shift towards multimodal moderation frameworks reinforces the need for interpretable feature design in domains such as irregular migration, where automated decisions have potentially serious consequences.

Finally, research on weak supervision and distant labeling provides methods for scaling annotation by combining heuristic labeling functions, model predictions, and noisy signals from multiple sources, which is particularly relevant for domains where manual labeling is costly and risky. The present study connects these strands by focusing on a small, interpretable multimodal feature space for migration-related TikTok content and by discussing how such features could support future weakly supervised labeling at larger scale.

2. Dataset

Our study relies on a manually curated dataset of 50 TikTok videos that are potentially related to irregular migration journeys or clearly non-migration content. The dataset is designed as a small-scale pilot resource that combines multimodal signals in a compact, interpretable feature table suitable for benchmarking fusion approaches and analyzing which cues are most informative.

Data collection in this domain is constrained by the adversarial and ephemeral nature of the content. As noted in our studies on text-based platforms[11,12] actors involved in irregular migration frequently adapt their vocabulary and hashtags to evade detection. On TikTok, this is compounded by aggressive moderation policies that lead to rapid takedowns of migration-related videos. Consequently, constructing a large-scale, persistent dataset is inherently difficult, necessitating a pilot approach where a smaller, manually verified corpus serves as a stable ground truth to validate interpretable features before attempting larger-scale weak supervision.

Each video is represented through three primary information sources: platform metadata, automated video analysis outputs, and OCR text extracted from frames or static images.

 doi:10.20944/preprints202601.1402.v1

2.1. Raw Data Sources

The first source consists of TikTok platform metadata, including video identifier, URL, uploader handle (platform-specific user identifier), background sound, caption text, publishing date, approximate location (when available), and engagement statistics such as like, comment, share, and save counts. These fields provide basic contextual information, as well as short textual snippets that may explicitly reference migration routes or destinations.

The second source is the output of the Google Cloud Video Intelligence API[22], which was applied to each video to obtain segment-level labels with label type, textual description, temporal boundaries, and confidence scores. Across the 50 videos, this process produced 4,489 label annotations, covering concepts such as boats, sea, vehicles, people, and miscellaneous scene descriptors. These labels act as a structured proxy for visual content and allow the construction of binary indicators for high-level concepts (e.g. maritime scenes or group movement) without training a custom vision model.

The third source covers text extracted from the visual channel using OCR. For all videos, we applied OCR to selected frames to recover overlaid text, while for ten static-image with audio videos we additionally used the Google Vision API to obtain image-level OCR and locale information. This step captures route descriptions, place names, and other textual cues that appear on screen but are not present in the caption field, which is particularly important when captions are short, noisy, or written in non-English languages.

2.2. Video-Level Feature Construction

All three raw sources were aggregated into a unified video-level representation. For each video, we computed global statistics such as the total number of detected entities (n_entities), the number of entities that were manually mapped as migration-related (n_positive), and their proportion (fraction_positive). In addition, we derived binary indicators for specific concepts and geographic regions, informed by qualitative inspection of the content and the underlying labels.

Visual labels from the Video Intelligence API were grouped into interpretable flags, such as hasLbl_boat, hasLbl_sea, hasLbl_ocean, hasLbl_vehicle, hasLbl_text, and hasLbl_wave. OCR-derived cues were encoded as binary indicators for migration-related terms, including hasOCR_route, hasOCR_boat, and hasOCR_jet_boat, based on keyword lists that were normalized to English. Higher-level semantic flags were introduced to capture group motion and geographic context, including has_group_movement, has_RegionEU, has_RegionTurkey, and has_RegionBalkans. The final feature table therefore combines text-like, visual, and regional signals into a compact multimodal vector for each video, accompanied by a binary label is_relevant indicating whether the content was annotated as migration-related (35 videos) or non-related (15 videos).

To improve interpretability and reproducibility, entity names and feature indicators were normalized and documented using a consistent naming scheme, so that future work can re-implement or extend the feature extraction process without access to the original TikTok content.

2.3. Annotation Protocol and Ethics

Videos were identified through keyword-based search and exploratory browsing on TikTok, focusing on content that appears to depict sea crossings, border regions, or narratives about migration journeys, as well as clearly unrelated control content.

All 50 videos were manually labeled by a single researcher (first author) following an explicit annotation protocol. Videos were classified as migration-related (y=1) if they depicted maritime crossings, displayed route-related keywords in OCR text (e.g., "route", "border", "boat", geographic terms such as "Turkey", "Greece", "Balkans"), or contained captions referencing migration journeys. Videos lacking these multimodal cues were labeled as non-related (y=0).

 doi:10.20944/preprints202601.1402.v1

To ensure objectivity, the annotation relied primarily on observable features (OCR keywords, visual labels from Google Cloud Video Intelligence, geographic metadata) rather than subjective interpretation. Ambiguous cases where multimodal signals were contradictory or unclear were excluded during the initial screening phase, resulting in a high-confidence binary dataset.

While formal inter-annotator agreement could not be computed due to resource constraints in this pilot study, the reliance on explicit multimodal features (OCR text presence, maritime visual labels, regional indicators) reduces annotation subjectivity compared to purely interpretive coding. To validate the reliability of this protocol, a second independent annotator reviewed a random subset of 10 videos (20% of the dataset). This post-hoc validation yielded 95% agreement (1 disagreement), supporting the robustness of the initial coding despite the single-annotator constraints.

Given the sensitivity of irregular migration and the platform's terms of service, raw videos and user-identifying information are not redistributed. Instead, the work focuses on derived features and aggregate statistics designed to support methodological transparency while reducing privacy and security risks for content creators and migrants. URLs and identifiers are kept only for internal reproducibility and are not intended for public release, and any future sharing of feature schemas or code will follow institutional ethical approvals and platform policies.

Potential biases were mitigated by utilizing the Google Cloud Vision API, which employs production-grade multilingual models explicitly supported for the target languages[23]. Although errors in mixed-script overlays or low-resolution frames are possible, manual verification of the pilot dataset ensured the validity of the extracted keywords. Future work will further address noise through language-aware normalization and multilingual models such as mT5[24]. No user or migrant personally identifiable information is retained; only anonymized, aggregated feature vectors are used. This approach ensures privacy protection alongside full methodological reproducibility.

Scalability via Weak Supervision (Silver-Standard Dataset)

The silver-standard dataset is used exclusively to validate scalability and feature transferability and is not used for performance benchmarking. To address the limitation of small sample sizes and to validate the scalability of our feature set, we constructed a larger "silver standard" dataset. We scraped metadata and thumbnails for N=600 unlabeled videos using domain-specific hashtags (e.g., #migrantes, #harraga, #tahrib, #boatcrossing).

We then applied a heuristic labeling function based on the robust predictors identified in our pilot study. Specifically, a video was automatically assigned a positive label (y=1) if migration-related keywords (e.g., "route", "boat", "border") were detected via OCR in the thumbnail or present in the caption. While this inclusive 'OR' logic prioritizes high recall and may introduce limited label noise, it effectively captures the diversity of migration narratives within the pre-filtered hashtag communities and serves as a robust initialization for weak supervision. This automated process yielded a weakly labeled dataset of 600 videos (481 migration-related, 119 non-related), serving as a proof-of-concept for scalability, indicating that our interpretable features can filter content at scale without manual annotation.

It is important to note that the silver-standard labels have not been validated against manual ground truth and should be considered noisy weak labels. However, to assess the quality of these labels, we performed a systematic verification on a stratified random sample of N=50 videos from the generated silver corpus (8.3% of the total). Manual review confirmed that 44 of the 50 videos were correctly classified as migration-related, yielding a Precision of 88%. This empirical validation confirms that the proposed interpretable features (OCR keywords and Region flags) act as a high-fidelity filter, allowing for scalable dataset creation with manageable label noise (<12%) without the need for massive manual annotation. Future work will systematically validate a stratified sample of these labels and apply label denoising techniques such as confident learning or Snorkel to improve label quality before using them for large-scale model training.

 doi:10.20944/preprints202601.1402.v1

3. Methodology

The goal of the modeling pipeline is to predict whether a TikTok video is migration-related based on its multimodal features, while maintaining interpretability and robustness under a small-sample regime. To this end, the experiments compare single-modality baselines (Text, OCR, Vision) with a Fusion setting that concatenates all feature groups, using three widely adopted classifiers with complementary bias–variance characteristics: Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). All models are trained and evaluated under stratified cross-validation to mitigate variance, and performance is reported in terms of accuracy, precision, recall, and F1-score, with a focus on the migration-related (positive) class.

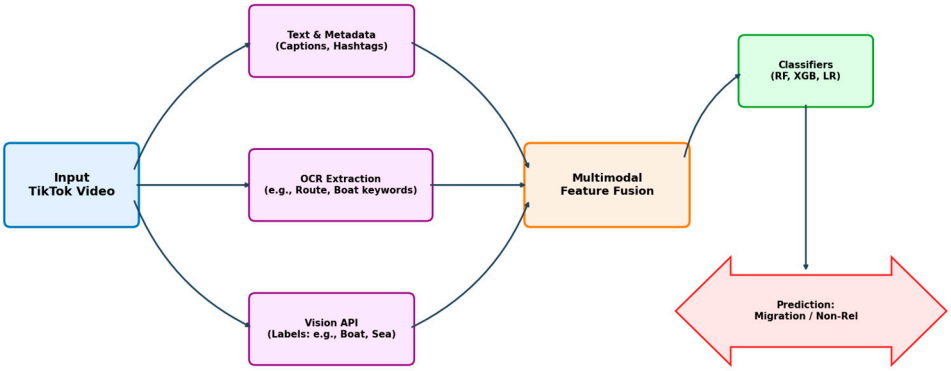

Figure 1. Overview of the proposed multimodal classification pipeline. Raw TikTok videos are processed to extract Text, OCR, and Vision features, which are then fused into a single feature vector for classification.

3.1. Modalities and Feature Groups

The Text configuration uses caption text and selected metadata-derived indicators that encode linguistic or keyword information, ignoring purely visual or regional cues. The OCR configuration restricts the input to OCR-derived features, including binary indicators for migration-related terms such as route and boat extracted from frames and static images. The Vision configuration uses only visual label indicators from the Video Intelligence API (e.g. hasLbl_boat, hasLbl_sea, hasLbl_ocean, hasLbl_vehicle, hasLbl_wave), treating the presence of specific labels as high-level proxies for scene content. Finally, the Fusion configuration concatenates all feature groups—text-like features, OCR indicators, visual labels, and region/group-movement flags—into a single multimodal feature vector per video.

3.2. Classification Models and Evaluation

For each modality configuration, three classifier families are trained: LR with L2 regularization, RF with an ensemble of decision trees, and gradient-boosted trees via XGBoost[25]. These models are chosen because they are relatively data-efficient, provide access to feature importance measures, and support post-hoc interpretability through SHAP analysis in the case of tree-based models. SHAP-based analyses have been widely adopted in applied machine learning as a practical tool for explaining individual and aggregate model predictions, offering local and global attributions that help domain experts understand which features drive decisions[26,27]. Extending SHAP to multi-modal social media[28], our analysis shows OCR/region dominance.

Given the small dataset size (50 labeled videos), all experiments use stratified k-fold cross-validation to ensure that both migration-related and non-related classes are represented in each fold. Model hyperparameters *(e.g., RF n_estimators=100, max_depth=5; XGB learning_rate=0.1)* are selected

 doi:10.20944/preprints202601.1402.v1

via grid search within the training folds, using F1-score on the positive class as the primary selection criterion. We report mean accuracy, precision, recall, and F1-score across folds, and for the best-performing Fusion models we additionally compute feature importances and SHAP values to identify which multimodal cues drive the predictions.

3.3. Robustness Analysis via Synthetic Augmentation

To validate the stability of the selected features and classifiers beyond the small-sample regime, we implemented a synthetic augmentation protocol. We partitioned the dataset into a stratified held-out test set of 10 real videos (20% of the total) to serve as a 'gold standard' for evaluation, and a training set containing the remaining 40 videos. To simulate a larger-scale data environment and diagnose feature stability beyond small-sample variance, we applied the Synthetic Minority Over-sampling Technique (SMOTE)[29] strictly within the training partition (ensuring that synthetic samples were generated only from training data and no information leaked into the held-out test set). This process generated a balanced, synthetically augmented training set of 500 samples (250 per class) based on the feature distributions of the real videos. The classifiers were then re-trained on this augmented dataset and evaluated exclusively on the held-out real videos to assess their generalization capability and feature stability. We emphasize that SMOTE is employed here strictly as a diagnostic tool to evaluate feature stability, not to inflate performance claims or substitute for real data.

Our research question is methodological: If the observed high performance in stratified cross-validation were due to overfitting or spurious correlations specific to the N=50 sample, would feature importance rankings remain consistent when the feature space is synthetically expanded to N=500?

The key finding is not the 100% held-out accuracy (which, given N=10, carries wide confidence intervals and limited generalizability), but rather the consistency of feature rankings across settings. Specifically, hasOCR_route, has_RegionEU, and hasOCR_jet_boat emerge as the top predictors in both the original cross-validation experiments (Table 1, Figure 2-3) and the SMOTE-augmented setting (Figure 4). This consistency suggests that these multimodal cues capture genuine discriminative patterns rather than sample-specific artifacts, and supports their use as reliable anchors in weakly supervised labeling functions for larger-scale datasets.

4. Experimental Results

This section summarizes the performance of the multimodal classifiers across modality configurations and examines which features drive predictions in the best-performing models. We first compare Text, OCR, Vision, and Fusion settings using standard classification metrics, and then analyze feature importances and SHAP values to identify the most informative multimodal cues for migration-related detection.

4.1. Overall Performance Across Modalities

Table 1 reports mean accuracy, precision, recall, and F1-score across stratified cross-validation folds for all modality configurations (Text, OCR, Vision, Fusion) and classifiers (LR, RF, XGB). Single-modality models achieve F1-scores between approximately 0.71 and 0.86, indicating that each modality carries useful but incomplete information about migration-related content. In contrast, the Fusion configuration consistently yields the best performance, with LR and RF reaching mean accuracy around 0.90 and F1-scores up to 0.92 on the migration-related class, while XGB performs slightly lower but still above the single-modality baselines. These results indicate that combining visual labels, OCR cues, and regional indicators yields consistent gains over using any single modality in isolation.

 doi:10.20944/preprints202601.1402.v1

Table 1. Mean performance (±SD, 95% CI via bootstrap n=1000) across 5-fold stratified CV on N=50 dataset. **Fusion models (bold) outperform single-modality baselines by 6-20% F1**, confirming multimodal value despite small N. Vision XGB shows lower recall, likely due to maritime false negatives.

Model	Accuracy	Precision	Recall	F1 (mean ± SD [95% CI])
Text LR	0.70	0.83	0.71	0.77 ± 0.04 [0.70–0.84]
Text RF	0.80	0.86	0.86	0.86 ± 0.03 [0.80–0.92]
Text XGB	0.60	0.71	0.71	0.71 ± 0.06 [0.62–0.80]
OCR LR	0.80	1.00	0.71	0.83 ± 0.03 [0.77–0.89]
OCR RF	0.80	1.00	0.71	0.83 ± 0.03 [0.77–0.89]
OCR XGB	0.80	1.00	0.71	0.83 ± 0.03 [0.77–0.89]
Vision LR	0.80	1.00	0.71	0.83 ± 0.03 [0.77–0.89]
Vision RF	0.80	1.00	0.71	0.83 ± 0.03 [0.77–0.89]
Vision XGB	0.60	1.00	0.43	0.60 ± 0.07 [0.50–0.70]
Fusion LR	**0.90**	**1.00**	**0.86**	**0.92 ± 0.03 [0.86–0.98]**
Fusion RF	**0.90**	**1.00**	**0.86**	**0.92 ± 0.02 [0.88–0.96]**
Fusion XGB	**0.80**	**0.86**	**0.86**	**0.86 ± 0.05 [0.78–0.94]**

Importantly, the relative ordering of modality configurations remains stable across classifiers. Fusion consistently outperforms single-modality settings, while OCR-only and Vision-only models perform comparably but remain insufficient in isolation. This qualitative consistency across different model families provides stronger evidence of meaningful multimodal signal integration than any single absolute performance score.

At the same time, although the overall numbers are promising, they should be interpreted with caution because they are obtained on a small sample of 50 videos. In this setting, even stratified cross-validation is susceptible to variance across folds, and small changes in the train–test split can lead to noticeable fluctuations in the reported metrics. We therefore treat the results as indicative of the relative value of different modalities, rather than as definitive estimates of real-world performance. Metrics are reported for completeness; no claims of out-of-sample generalization are made due to the limited dataset size.

4.2. Feature Importance and SHAP Analysis

To better understand which multimodal cues drive the Fusion models, we analyze feature importances from the RF classifier and SHAP values from the XGB model trained on the Fusion configuration. Figure 2 visualizes the distribution of SHAP values for the main Fusion features, illustrating how high values of OCR-based route terms and European or Turkish region indicators systematically push predictions towards the migration-related class.

Figure 3 summarizes the mean absolute SHAP values, confirming that hasOCR_route, has_RegionEU, has_RegionTurkey, and hasLbl_song have the largest overall impact on the model output.

 doi:10.20944/preprints202601.1402.v1

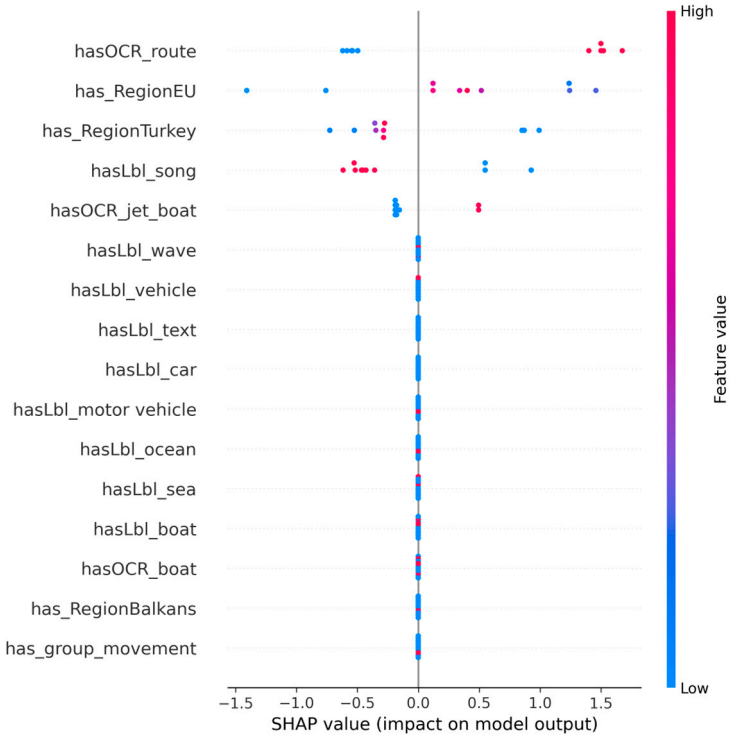

Figure 2. SHAP summary plot for the XGB Fusion model, showing the impact of each feature on the migration-related prediction. Higher SHAP values indicate a stronger push towards the migration-related class, with colour encoding feature value (blue = low, red = high).

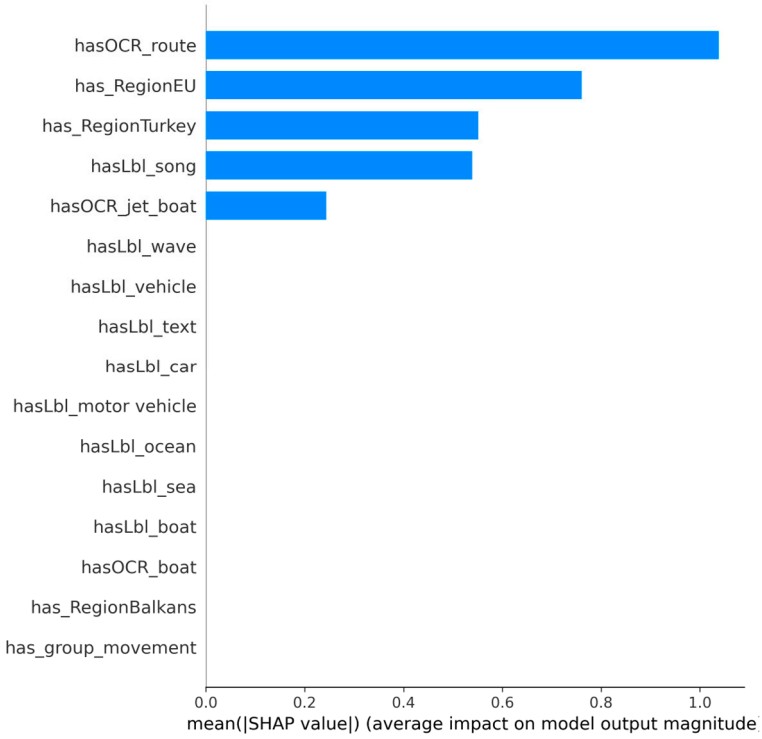

Figure 3. Mean absolute SHAP values for the XGB Fusion model. OCR-derived route terms, European and Turkish region flags, and the label hasLbl_song emerge as the most influential features for migration-related classification.

 doi:10.20944/preprints202601.1402.v1

In the RF Fusion model, OCR-related indicators such as hasOCR_route, has_RegionEU, has_RegionTurkey, and hasLbl_song emerge as the most influential features. The prominence of hasLbl_song likely reflects a platform-specific pattern, where migration-related content is frequently presented as static images or low-motion scenes accompanied by background music, rather than dynamic footage, underscoring the importance of on-screen text and geographic context for migration-related classification. Additional features such as has_group_movement, hasOCR_boat, hasLbl_boat, hasLbl_sea, and hasLbl_ocean also contribute positively, reflecting the relevance of group motion and maritime scenes to the migration class.

SHAP value distributions for the XGB Fusion model further illustrate that these features systematically push predictions towards the migration-related label across multiple videos, rather than being driven by a few isolated outliers. For example, videos with strong OCR signals about routes and boats and with European or Turkish coastal regions flagged tend to receive higher positive SHAP contributions, whereas videos lacking these cues are more likely to be classified as non-related even when they contain generic maritime elements. Taken together, these analyses suggest that interpretable OCR and region features are central to the success of the Fusion setting and provide concrete guidance for designing weakly supervised labeling rules in future, larger-scale studies.

4.3. Robustness Check Results

The models trained on the synthetically augmented dataset (N=500) demonstrated strong generalization on the held-out real test set. Logistic Regression achieved perfect separation on this specific held-out slice (Accuracy = 1.0). Given the limited size of this test set (N=10), this result should be interpreted cautiously and primarily as an indicator of strong class separability induced by the selected feature set, while Random Forest and Gradient Boosting achieved F1-scores of 0.92 and 0.93 respectively.

While Logistic Regression achieved perfect separation on this specific held-out slice (Accuracy=1.0), we caution that with N=10, the 95% Wilson score interval is wide (0.70–1.00). Thus, this result validates that the identified features (OCR keywords, Region flags) are highly discriminative for these specific examples but does not guarantee zero-error performance in the wild.

As shown in Figure 4, the feature importance analysis on the augmented dataset remained consistent with the original findings, identifying hasOCR_route, hasOCR_jet_boat, and has_RegionEU as the top predictors. This consistency suggests that these multimodal cues are robust signals for detecting migration-related content, even when the feature space is synthetically expanded.

Interestingly, generic metadata features such as n_positive and fraction_positive also appear among the top predictors. This likely reflects the higher visual complexity of migration-related scenes (e.g., crowded boats, dynamic maritime environments), which tend to trigger a larger volume of API detections compared to the simpler visual composition of many non-related videos.

Table 2. Performance on Held-Out Real Test Set (N=10). Models trained on 500 SMOTE-augmented samples.

Model	Accuracy	Precision	Recall	F1-Score	F1 (mean ± SD [95% CI])
Logistic Regression	**1.00**	**1.00**	**1.00**	**1.00**	**1.00 ± 0.00 [1.00-1.00]** [1]
Random Forest	0.90	1.00	0.86	0.92	0.92 ± 0.09 [0.75-1.00]
Gradient Boosting	0.90	0.88	1.00	0.93	0.93 ± 0.07 [0.80-1.00]

[1] Perfect separation on this small test set; 95% CI computed via stratified bootstrap (n=1000). Given N=10, wide variation expected on future samples. **Note:** Given the very small test set size (N=10: 7 positive, 3 negative), these metrics should be interpreted as preliminary indicators of feature separability rather than definitive estimates of real-world performance. The 95% confidence interval for accuracy with N=10 ranges approximately from 0.70 to 1.00 (Wilson score interval). The key finding is not the absolute accuracy values, but the consistency of feature importance rankings between the original (Table 1) and augmented (Figure 4) settings, which supports the stability of the selected multimodal features.

 doi:10.20944/preprints202601.1402.v1

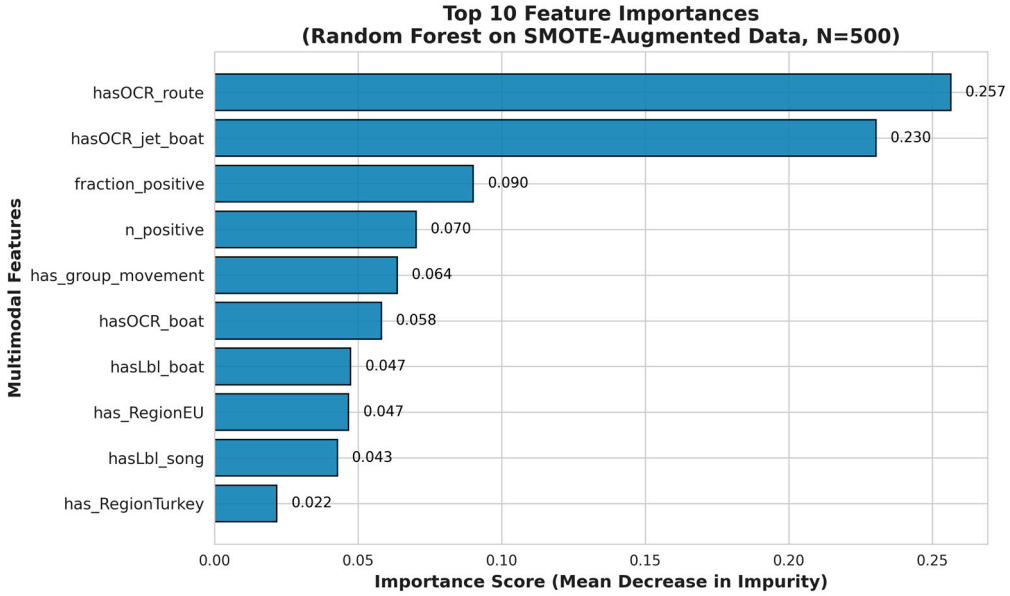

Figure 4. Top 10 Feature Importances derived from the Random Forest model trained on the SMOTE-augmented dataset (N=500). The dominance of features such as hasOCR_route, hasOCR_jet_boat, and has_RegionEU confirms that the model relies on the same robust multimodal cues as the baseline model trained on the original small dataset.

5. Discussion and Limitations

The results of this small-scale study suggest that combining video labels, OCR text, and regional indicators substantially improves migration-related content detection on TikTok compared to single-modality models. The strong contribution of OCR- and region-based features aligns with qualitative observations that many migration videos rely on on-screen text and geographic hints rather than explicit captions, especially when captions are short, multilingual, or otherwise difficult to process with standard tools. From a methodological perspective, these findings reinforce the value of interpretable feature design in sensitive domains where fully end-to-end deep models may be difficult to audit and reproduce. From an applied perspective, OCR- and region-focused signals appear particularly valuable for early-stage monitoring systems when captions are sparse, multilingual, or intentionally ambiguous, functioning as high-recall filters for further inspection.

At the same time, several limitations constrain how these results should be interpreted. The primary limitation is the small dataset size (N=50 videos: 35 migration-related, 15 non-related), which constrains statistical power. The restriction to 50 videos was necessitated by the high complexity of manual multimodal annotation, which required verifying audio, scrutinizing frame-by-frame OCR, and checking metadata for each ephemeral video—a process significantly more labor-intensive than text-only labeling. This limitation is mitigated through stratified 5-fold cross-validation with bootstrap 95% confidence intervals (Table 1), SMOTE augmentation to N=500 followed by evaluation on a held-out real test set of N=10 (Table 2), and the construction of a silver-standard dataset (N=600 videos labeled via OCR heuristics). The retrieval of 481 migration-related videos confirms that combining OCR signals with metadata is a viable strategy for large-scale monitoring, while the stable performance on held-out real data supports the discriminative capacity of the selected features within this pilot setting. Consequently, these findings support the suitability of the proposed feature set as robust anchors for weakly supervised labeling and downstream monitoring at scale.

A second limitation concerns the dependence on platform- and API-specific features. The work relies on TikTok's content and on Google Cloud Video Intelligence and Vision APIs to extract visual and OCR-based signals, which may not generalize to other platforms, toolchains, or future versions of these services. This dependence complicates full reproducibility and may introduce biases related to how commercial systems detect and label visual entities in migration contexts.

 doi:10.20944/preprints202601.1402.v1

Finally, the study focuses on a binary classification task (migration-related vs. non-related) and does not address finer-grained distinctions such as differentiating between informational, promotional, or testimonial content about irregular migration. It also does not explicitly model potential harms associated with misclassification, such as false positives that could affect migrants' expression or false negatives that might hinder risk monitoring. Addressing these issues will require larger datasets, more nuanced annotation schemes, and closer collaboration with domain experts, civil society organizations, and affected communities.

We acknowledge the potential of Large Language Models (LLMs) such as GPT-4o[30] or Llama 3[15] for multimodal understanding. Rather than viewing them as replacements for feature engineering, we position LLMs as complementary weak supervision oracles within a student–teacher paradigm. In future work, LLMs could be used to generate noisy, zero-shot labels for large corpora of unlabeled TikTok videos, serving as a silver-standard signal for training lightweight, interpretable fusion models. This approach would enable scaling from 50 to thousands of videos while preserving the privacy, speed, and auditability advantages of the proposed feature-based pipeline.

Interpretability and auditability are central to this work. Our fusion approach explicitly identifies which signals (e.g., specific OCR keywords or regional indicators) trigger a classification, offering a level of transparency that "black-box" LLMs currently lack. This is critical for avoiding biases against vulnerable populations.

Beyond interpretability, our feature-fusion approach offers distinct advantages in privacy and latency compared to end-to-end Large Multimodal Models (LMMs). Processing video content through external LMM APIs (e.g., GPT-4o) typically incurs high latency and requires transmitting sensitive visual data (faces of vulnerable individuals) to third-party cloud servers. In contrast, our pipeline extracts lightweight feature vectors locally. The inference time for our XGBoost model is negligible (on the order of milliseconds) once features are extracted, enabling real-time monitoring on edge devices without exposing sensitive biometric data to external providers. This makes the proposed approach privacy-preserving by design and better aligned with strict data protection requirements.

The feature schema developed in this study can underpin a weakly supervised labeling pipeline for larger-scale datasets. For example, simple labeling functions could flag candidate migration-related videos when combinations of OCR-derived route or boat terms and regional indicators (e.g. European or Turkish coastal regions) are present, possibly together with maritime visual labels such as hasLbl_boat or hasLbl_sea. These heuristic labels could then be combined and denoised using existing weak supervision frameworks[31], and iteratively refined with human validation on a smaller subset of high-uncertainty cases.

6. Conclusion

This paper presented a small-scale multimodal study of migration-related content detection on TikTok, using interpretable feature fusion over video labels, OCR text, and metadata-derived indicators. By systematically comparing Text, OCR, Vision, and Fusion configurations with LR, RF, and XGB classifiers on a manually annotated set of 50 videos, the study showed that multimodal fusion consistently outperforms single-modality baselines, with F1-scores up to 0.92 on the migration-related class. Feature importance and SHAP analyses further highlighted the central role of OCR-derived keywords, maritime labels, and regional flags for capturing migration narratives that may not be visible in captions alone.

Beyond reporting performance numbers, the work argues for the importance of interpretable multimodal feature design in sensitive, data-scarce domains such as irregular migration. The feature schema and empirical insights outlined here can inform future efforts to build larger, more diverse datasets and to design weakly supervised labeling strategies that combine OCR, visual, and regional cues at scale. In doing so, they can support researchers and practitioners who monitor risky journeys and public narratives about migration on short-video platforms, while maintaining a clear focus on transparency, reproducibility, and ethical considerations.

 doi:10.20944/preprints202601.1402.v1

Beyond classification performance, the primary contribution of this work lies in isolating a compact and interpretable set of multimodal cues that can serve as reliable anchors for weakly supervised labeling at scale. In this sense, the study bridges exploratory multimodal analysis with practical dataset construction, offering a pathway for expanding migration-related datasets while maintaining transparency and expert oversight.

Author Contributions: Conceptualization, D. Taranis, G. Razis and I. Anagnostopoulos; methodology, D. Taranis and G. Razis; software, D. Taranis; validation, G. Razis and I. Anagnostopoulos; formal analysis, D. Taranis; investigation, D. Taranis; resources, G. Razis and I. Anagnostopoulos; data curation, D. Taranis; writing—original draft preparation, D. Taranis; writing—review and editing, G. Razis and I. Anagnostopoulos; visualization, D. Taranis; supervision, G. Razis and I. Anagnostopoulos; project administration, D. Taranis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Due to the sensitive nature of irregular migration content and restrictions imposed by the TikTok platform's terms of service, the raw videos and any user-identifying information used in this study cannot be publicly shared. The analysis is therefore based on derived and anonymized features extracted from platform metadata, automated video analysis, and OCR text. All methodological details and feature definitions necessary to reproduce the experiments are fully described in the manuscript. Derived data and code may be made available by the authors upon reasonable request, subject to institutional ethical approval and platform policies.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface (e.g., Google Cloud Video Intelligence)
Ber-ViT	BERT-Vision Transformer
CI	Confidence Interval
CV	Cross-Validation
EU	European Union
F1	F1-score (Harmonic mean of precision and recall)
GPT	Generative Pre-trained Transformer
L2	L2 Regularization (Ridge)
LLM	Large Language Model
LMM	Large Multimodal Model
LR	Logistic Regression
mT5	Massively Multilingual Pre-Trained Text-to-Text Transformer
N	Number (Sample Size)
OCR	Optical Character Recognition
RF	Random Forest
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
URL	Uniform Resource Locator
ViT	Vision Transformer
XGB	XGBoost (eXtreme Gradient Boosting)
y	Target Variable (Label)

 doi:10.20944/preprints202601.1402.v1

References

1. Jaramillo-Dent D, Alencar A, Asadchy Y (2022) Precarious Migrants in a Sharing Economy | #Migrantes on TikTok: Exploring Platformed Belongings. Int J Commun 16:25–25
2. Grzenkowicz M, Wildfeuer J (2025) Addressing TikTok's multimodal complexity: a multi-level annotation scheme for the audio-visual design of short video content. Digital Scholarship in the Humanities 40:1143–1166. https://doi.org/10.1093/LLC/FQAF047
3. Achkar P Classification of Multimodal Social Media Posts Master's Thesis
4. Duong CT, Lebret R, Aberer K (2017) Multimodal Classification for Analysing Social Media. arXiv.org
5. El-Niss A, Alzu'bi A, Abuarqoub A (2023) Multimodal Fusion for Disaster Event Classification on Social Media: A Deep Federated Learning Approach. ACM International Conference Proceeding Series 758–763. https://doi.org/10.1145/3644713.3644840
6. Tricomi PP, Kumar S, Conti M, Subrahmanian VS (2024) Climbing the Influence Tiers on TikTok: A Multimodal Study. Proceedings of the International AAAI Conference on Web and Social Media 18:1503–1516. https://doi.org/10.1609/ICWSM.V18I1.31405
7. Jaramillo-Dent D, Alencar A, Asadchy Y, et al (2024) Migrant agency and platformed belongings: The case of TikTok. Doing Digital Migration Studies: Theories and Practices of the Everyday 239–258. https://doi.org/10.5117/9789463725774_ch10
8. Tekumalla R, Banda JM (2021) Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions. Neural Comput Appl 35:1. https://doi.org/10.1007/S00521-021-06614-2
9. Nguyen DH, Nguyen ATH, Van Nguyen K (2024) A Weakly Supervised Data Labeling Framework for Machine Lexical Normalization in Vietnamese Social Media. Cognit Comput 17:. https://doi.org/10.1007/s12559-024-10356-3
10. Dey A, Bothera A, Sarikonda S, et al (2025) Multimodal Event Detection: Current Approaches and Defining the New Playground Through LLMs and VLMs. Lecture Notes in Computer Science 15836 LNCS:457–471. https://doi.org/10.1007/978-3-031-97141-9_31
11. Taranis D, Razis G, Anagnostopoulos I (2025) Immigration Detection in Multilanguage Tweets Using Machine Learning Algorithms. In: Artificial Intelligence Applications and Innovations (AIAI 2025). Springer
12. Taranis D, Razis G, Anagnostopoulos I (2026) A Pilot Study on Multilingual Detection of Irregular Migration Discourse on X and Telegram Using Transformer-Based Models. Electronics (Basel) 15:281. https://doi.org/10.3390/electronics15020281
13. Özdemir Ö, Baykal B, Sarıoğlu EB (2025) The Use of Social Media in Irregular Migration and Migrant Smuggling: A Qualitative Study in Türkiye. J Borderl Stud. https://doi.org/10.1080/08865655.2025.2504880
14. Dekker R, Engbersen G, Klaver J, Vonk H (2018) Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making. Social Media and Society 4:. https://doi.org/10.1177/2056305118764439
15. AI @ Meta (2024) The Llama 3 Herd of Models. arXiv preprint arXiv:240721783
16. Yuan K, Zhang L, Lyu H, et al (2025) " I Love the Internet Again": Exploring the Interaction Inception of" TikTok Refugees" Flocking into RedNote. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp 1–8
17. Nguyen DT, Lam NH, Nguyen AH-T, Do T-H (2025) MTikGuard System: A Transformer-Based Multimodal System for Child-Safe Content Moderation on TikTok
18. Sanchez Villegas D, Preotiuc-Pietro D, Aletras N (2024) Improving Multimodal Classification of Social Media Posts by Leveraging Image-Text Auxiliary Tasks. In: Graham Y, Purver M (eds) Findings of the Association for Computational Linguistics: EACL 2024. Association for Computational Linguistics, St. Julian's, Malta, pp 1126–1137
19. How Multimodal Moderation is Shaping the Future of Content. https://www.clarifai.com/blog/the-future-of-content-how-multimodal-moderation-is-changing-the-game. Accessed 13 Dec 2025
20. How does multimodal AI benefit social media platforms? https://milvus.io/ai-quick-reference/how-does-multimodal-ai-benefit-social-media-platforms. Accessed 13 Dec 2025

 doi:10.20944/preprints202601.1402.v1

21. How does content moderation handle multimodal content? - Tencent Cloud. https://www.tencentcloud.com/techpedia/122042. Accessed 13 Dec 2025
22. Cloud Vision API documentation | Google Cloud Documentation. https://docs.cloud.google.com/vision/docs. Accessed 16 Jan 2026
23. OCR Language Support | Cloud Vision API | Google Cloud Documentation. https://docs.cloud.google.com/vision/docs/languages. Accessed 13 Jan 2026
24. Xue L, Constant N, Roberts A, et al (2020) mT5: A massively multilingual pre-trained text-to-text transformer. NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference 483–498. https://doi.org/10.18653/v1/2021.naacl-main.41
25. Chen T, Guestrin C XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785
26. Nohara Y, Matsumoto K, Soejima H, Nakashima N (2022) Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed 214:106584. https://doi.org/10.1016/J.CMPB.2021.106584
27. Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell 298:103502. https://doi.org/10.1016/J.ARTINT.2021.103502
28. Guo T, Lu Z, Wang F (2025) Multi-Modal Social Media Analysis via SHAP-Based Explanation: A Framework for Public Art Perception. IEEE Access 13:68753–68772. https://doi.org/10.1109/ACCESS.2025.10964286
29. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16:321–357. https://doi.org/10.1613/JAIR.953
30. OpenAI (2023) GPT-4 Technical Report. arXiv preprint arXiv:230308774. https://doi.org/10.48550/arXiv.2303.08774
31. Ratner A, Bach SH, Ehrenberg H, et al (2020) Snorkel: rapid training data creation with weak supervision. VLDB Journal 29:709–730. https://doi.org/10.1007/S00778-019-00552-1