SAGE: Scalable Automatic Gating Ensemble for Confident Negative Harvesting in Fraud Detection

Sudheer Tubati
sudheet@amazon.com
Amazon
Seattle, USA

Amit Goyal
goyalam@amazon.com
Amazon
San Francisco, USA

Abstract

Music streaming fraud, where bad actors artificially inflate stream counts to manipulate chart rankings and royalty payments, poses a significant threat to streaming services and legitimate content creators. Traditional fraud detection approaches struggle with a critical challenge: many legitimate edge cases, including super-fans and sleep-music sessions, exhibit activity patterns that closely mimic those of coordinated fraud. We present SAGE, a novel counterfactual-aware negative harvesting approach that combines SimHash-based stratified sampling with a modular gating ensemble for confident negative identification from unlabeled data. Our ensemble architecture employs pluggable statistical gates (currently instantiated with Mahalanobis distance and k-NN density) with configurable voting thresholds enabling adaptive precision-recall trade-offs. This addresses the representation bias problem in Positive-Unlabeled learning by ensuring comprehensive coverage of rare behavioral cohorts through floor-constrained sampling. Evaluation demonstrates strong precision and recall on held-out data. The approach generalizes across fraud detection domains, achieving strong performance on both customer-level and artist-level fraud without modification to the core methodology.

CCS Concepts

• **Computing methodologies** → **Semi-supervised learning settings**; **Anomaly detection**; *Classification and regression trees*; Cluster analysis; • **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Information systems** → *Multimedia streaming*.

Keywords

Music Stream Fraud, Fraud Detection, Bot Farms, Streaming Manipulation

ACM Reference Format:
Sudheer Tubati and Amit Goyal. 2026. SAGE: Scalable Automatic Gating Ensemble for Confident Negative Harvesting in Fraud Detection. In *The Nineteenth ACM International Conference on Web Search and Data Mining (WSDM Companion '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3779211.3793166

WSDM Companion '26, Boise, ID, USA

ACM ISBN 979-8-4007-2358-2/2026/02
https://doi.org/10.1145/3779211.3793166

1 Introduction

The music streaming industry has experienced unprecedented growth, with global recorded music revenue reaching $29.6 billion in 2024, where streaming services account for 69% ($20.4 billion) of this total [12]. In the U.S., streaming represents 84% of the market, generating over $14 billion in revenue with nearly 97 million paid subscriptions [26]. However, this financial success has attracted sophisticated fraud schemes, with investigations showing that 1-3% of streams on major services may be fraudulent [6].

In 2024, the first-ever criminal prosecution for music streaming fraud charged a North Carolina musician with running a 7-year scheme using AI-generated songs and bot networks that generated over $1.2 million in annual royalties [28]. These incidents underscore how streaming manipulation combines technological sophistication with scale to exploit royalty payment systems.

The industry has responded with increased collaboration and technological deployment. In late 2023, a coalition including Spotify, Amazon Music, and Universal Music Group formed the "Music Fights Fraud" task force [23]. Music streaming services are deploying advanced AI-based detection systems at scale, with solutions processing over 2 trillion streams in 2023 [22], reflecting growing recognition that robust fraud detection is essential for protecting artist revenues and maintaining ecosystem integrity.

1.1 Related Work

PU Learning. Positive-Unlabeled learning [2, 9, 18] addresses scenarios where only positive examples are labeled. Recent advances include nnPU [16] and its variants [5], which use risk estimators to handle label noise during training. However, these methods require iterating over the entire unlabeled population, which becomes computationally prohibitive at global streaming service scale.

LSH and Sampling. SimHash [4] preserves similarity through random projections, widely used for near-duplicate detection [20] and nearest neighbor search [13]. We adapt SimHash for behavioral stratification to ensure rare cohorts are represented, a novel application to address representation bias in fraud detection. Our floor-constrained sampling ensures minimum representation per behavioral stratum, preventing false positives on edge cases.

Ensemble Methods and Statistical Gates. Ensemble approaches combine multiple models or filters to improve robustness [8]. Traditional ensembles aggregate predictions from multiple classifiers (bagging, boosting, stacking), while our gating ensemble operates at the data curation stage, filtering samples for training set construction. Mahalanobis distance [19] with Ledoit-Wolf shrinkage [17] provides robust multivariate outlier detection, while k-NN-based methods [3, 25] capture local density. Our contribution is a modular gating ensemble architecture where pluggable statistical gates

vote on sample confidence for PU learning, with configurable voting thresholds enabling adaptive precision-recall trade-offs. Unlike prediction ensembles, our gates ensure samples pass both global statistical and local density checks before inclusion in training data.

Fraud Detection. Streaming fraud detection remains largely unexplored. Esmaeilzadeh et al. [10] use heuristic labeling for video streaming, while Sejr et al. [11] apply outlier detection to music data. Related work includes bot detection [21] and financial fraud [7, 24]. However, none address the combination of counterfactual-aware negative harvesting, global traffic scale, extreme class imbalance, and absence of systematic negative labels. These are the core challenges SAGE is designed to solve.

2 Data and Features

Our fraud detection system operates on customer-level behavioral data aggregated from streaming interactions over defined observation windows, formulated as binary classification distinguishing Fraud (manipulated activity) from Non-Fraud (legitimate engagement). The feature engineering process evolved through multiple iterations, guided by domain expertise and data-driven insights, converging on a representation balancing predictive power with real-time deployment constraints.

The feature space captures three critical dimensions of streaming behavior. First, we compute temporal consistency patterns through variance metrics across hourly, daily, and weekly time grains, where low variance typically indicates bot-like uniform activity while high variance suggests organic human usage. Second, we extract behavioral diversity signals including entropy and standard deviation over categorical attributes such as device types and content selection sources, where high entropy reflects varied human-like interactions and low entropy may indicate scripted behavior. Third, we compute short-term behavioral trends over a trailing observation window that prove particularly effective at distinguishing emerging fraud patterns from legitimate edge cases like super-fans or ambient listeners. This design prioritizes interpretability and computational efficiency, avoiding complex sequence models in favor of explainable statistical features that align with risk investigator intuition.

Our labeling strategy evolved significantly throughout the project life cycle. Initially, we bootstrapped training using heuristic labels derived from domain expertise and risk specialist knowledge due to the scarcity of high-quality annotated data. These heuristics partitioned the customer base into categorized and uncategorized segments, revealing extreme class imbalance with fraud cases representing approximately 1% of the labeled population, characteristic of fraud detection problems. As our pipeline matured and we developed custom investigation tools, we transitioned to human-labeled ground truth collected through systematic manual review. This progression from heuristic to human-annotated labels enabled increasingly sophisticated modeling approaches and represents a critical evolution in our detection capabilities. The final feature set, refined through model importance scores, comprises multiple dimensions optimized for production-scale fraud detection.

3 Approach

3.1 Evolution of Prior Work

Our research progressed through multiple modeling paradigms, each addressing limitations of its predecessor. We began with unsupervised anomaly detection using Isolation Forest and Variational Autoencoders [15], which surfaced potential fraud through anomaly scores but lacked the precision required for production deployment. We then explored semi-supervised approaches including random undersampling of the majority class and student-teacher self-learning frameworks [1, 29], which initially improved precision but suffered from error amplification in iterative self-labeling, leading to precision drops, along with label noise and incomplete population coverage. Cluster-based methods using K-Means for intelligent undersampling [30] showed further gains by preserving population structure.

However, a critical blind spot persisted across all approaches: underrepresentation of low-tenure customers and certain device cohorts in training data created systematic bias, where models flagged these segments as fraud simply due to insufficient exposure to their normal behavioral patterns. This counterfactual problem (the absence of legitimate edge cases that resemble fraud) motivated SAGE.

3.2 Proposed Approach

The core challenge in streaming fraud detection lies not in identifying fraud, but in comprehensively modeling legitimate behavior at scale. Traditional fraud-first approaches struggle with this, leading to elevated false positives. We adopt a "understand the haystack before looking for needles" philosophy, addressing the counterfactual problem through SAGE: a modular gating ensemble architecture combining locality-sensitive hashing for behavioral stratification with pluggable statistical gates for confident negative harvesting.

3.2.1 SimHash-Based Stratified Sampling: We employ SimHash [4], a locality-sensitive hashing technique, to map high-dimensional customer behavior into binary signatures where behaviorally similar customers receive similar hash codes. This enables efficient stratification of millions of daily customers into thousands of behavioral buckets, ensuring proportional representation across the entire behavioral spectrum. Unlike random sampling, which underrepresents rare cohorts, our approach applies floor constraints (minimum samples per bucket) to guarantee coverage of edge cases such as super-fans and functional music listeners. SimHash-based stratification operates in O(n) time, preserving population structure while solving the representation bias problem that plagued earlier iterations. We tuned hyperparameters for this component (bucket count and floor threshold) independently to optimize behavioral coverage rather than through end-to-end grid search.

3.2.2 Gating Ensemble for Confidence Filtering: To harvest high-confidence negatives from the unlabeled population, we introduce a modular ensemble of statistical gates where each gate independently assesses sample confidence. The ensemble employs a configurable voting mechanism: samples are accepted based on how many gates they pass, allowing adaptive precision-recall trade-offs. For high-precision scenarios, we require unanimous voting (all gates

must pass). For higher recall, we can relax to majority voting or k-out-of-n thresholds. This flexibility enables us to tune the operating point based on deployment constraints without retraining.

This ensemble architecture is extensible; additional gates such as Isolation Forest, Local Outlier Factor, or domain-specific heuristics can be plugged in without modifying the core framework. Our current instantiation employs two complementary gates fitted on the labeled fraud distribution. The first gate uses Mahalanobis distance [19] with Ledoit-Wolf covariance shrinkage [17] to measure global statistical distance from the fraud distribution's center, accounting for feature correlations and variance. Samples exceeding a calibrated distance threshold (far from fraud in global feature space) pass this gate. The second gate employs k-NN density estimation [3] to capture local structure. A sample might be globally distant from fraud but locally surrounded by fraud instances, or vice versa.

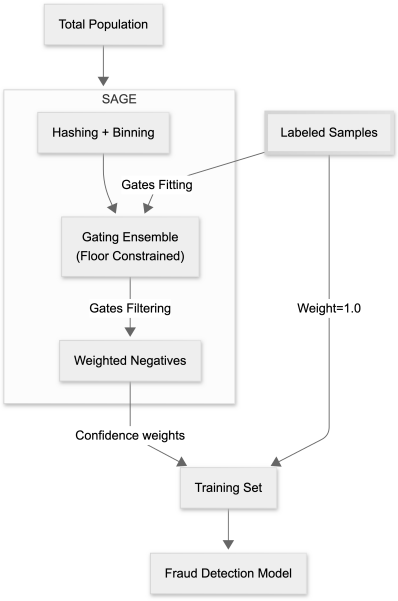

Figure 1: SAGE - SimHash stratification with floor-constrained gating ensemble to harvest confident negative samples with weights

For our current model, we use unanimous voting (both gates must pass) to maximize precision in confident negative harvesting, effectively filtering contamination through complementary global and local perspectives. The dual-gate design addresses a statistical limitation: global distance alone would accept boundary samples near the fraud distribution's edge, while local density alone would reject legitimate outliers surrounded by sparse fraud instances. The combination ensures samples are both globally distant and locally dissimilar to fraud. We assign harvested negatives confidence-based weights derived from their gate scores, with samples passing gates by larger margins receiving higher weights in training.

Threshold tuning follows a systematic procedure: we sweep candidate thresholds for each gate independently on held-out validation data, measuring contamination rates (samples incorrectly harvested as negatives). Contamination is assessed using distance-based metrics and simple regression models. Thresholds are selected to minimize contamination while maximizing negative sample yield. The dual-gate combination with unanimous voting is then validated on a separate test set to confirm low contamination is maintained.

To validate the effectiveness of our gating ensemble design, we conducted ablation experiments comparing different combinations of SimHash stratification and statistical gates. Figure 2 shows relative precision-recall comparison demonstrating that the combination of SimHash with dual gates (Mahalanobis + k-NN) with bin floors achieves superior performance compared to individual components or alternative configurations, confirming the complementary nature of our ensemble architecture. Figure 1 illustrates the complete pipeline from SimHash stratification through the gating ensemble to final training set construction.

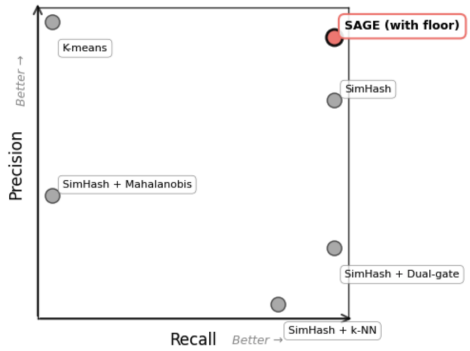

Figure 2: Precision-recall chart for ablation study comparing SimHash stratification and gating combinations

3.2.3 Multi-Class Formulation and Training: While the core problem is binary classification (Fraud vs. Non-Fraud), human investigations revealed cases where investigators were uncertain about the fraud label. To maintain high precision and avoid forcing uncertain cases into binary labels, we introduce a third class called Suspicious and formulate the problem as multi-class classification. The foundation dataset comprises human-labeled fraud and suspicious cases from risk operations investigations, along with a small set of manually verified non-fraud samples. However, this labeled non-fraud set is insufficient to represent the full spectrum of legitimate behavior. SAGE's SimHash-gating ensemble approach addresses this by harvesting confident negatives from the unlabeled population, expanding the non-fraud set by orders of magnitude while maintaining comprehensive behavioral coverage. We train a LightGBM [14] classifier on this hybrid dataset, incorporating behavioral features (temporal variance, entropy, standard deviation), ambient listening patterns, and customer attributes. The primary focus remains on the Fraud class for production decisions, while the Suspicious class provides a buffer for uncertain cases requiring manual review. This counterfactual-aware training ensures the model has seen diverse legitimate behaviors before encountering similar patterns in production, reducing false positives on edge cases.

The model operates with a two-tier decision framework. High-confidence fraud predictions trigger immediate automated annotation, while lower-confidence scores route to manual review queues. This hybrid approach balances automation with human oversight [27], achieving high precision with operationally useful recall for production deployment at scale.

4 Results

We evaluated our approach against multiple baseline methods, progressing from unsupervised anomaly detection through semi-supervised and self-learning approaches. Table 1 summarizes performance improvements relative to an Isolation Forest baseline. While prior methods achieved substantial gains, all shared a critical limitation: systematic under-representation of low-tenure customers and certain device cohorts in training data, leading to elevated false positive rates on these edge case segments.

Note that nnPU [16] is not included in baseline comparisons. We trained nnPU on a random undersample of unlabeled data, but this is not comparable to other approaches in this paper as it did not leverage the full unlabeled population. Training nnPU on the full unlabeled population is computationally infeasible (millions of daily customers), which is where scale becomes a critical differentiator for SAGE.

Table 1: Relative Performance (percentage points) over baseline

Method	Δ Precision (pp)	Δ Recall (pp)	Δ F1 (pp)
Isolation Forest	baseline	baseline	baseline
Var. Auto Enc.	+55.8	+10.1	+19.2
Random undersampling	+73.8	**+82.2**	+78.6
Student-Teacher	**+82.0**	+22.3	+36.3
Clustering (K-Means)	+80.5	+75.0	+78.0
SAGE (proposed)	**+81.9**	**+87.2**	**+85.2**

SAGE addresses the counterfactual problem directly and achieves the strongest overall performance, with balanced precision and recall improvements of +81.9pp and +87.2pp respectively. Evaluation on held-out data demonstrates performance across diverse customer segments, including previously problematic edge cases.

5 Conclusion

We present SAGE, a scalable approach for confident negative harvesting in Positive-Unlabeled learning that addresses the counterfactual problem, the absence of representative negative examples resembling positive cases. The core challenge in streaming fraud detection lies not in identifying fraud, but in comprehensively modeling legitimate behavior at scale. SAGE adopts a "understand the haystack before looking for needles" philosophy, combining SimHash-based stratified sampling with a modular gating ensemble featuring configurable voting thresholds. This enables adaptive precision-recall trade-offs without retraining. Our current model achieves +81.9pp precision and +87.2pp recall improvements over baseline, demonstrating that comprehensive legitimate behavior modeling outperforms fraud-first approaches.

The modular architecture generalizes across fraud detection domains and applies to any classification problem with scarce positive labels, absent negative labels and edge cases resembling the target class in domains like financial fraud, bot detection, spam filtering, and cybersecurity. SAGE's paradigm shift from detecting anomalies to understanding the full behavioral spectrum offers a principled path for robust classification at scale in challenging real-world scenarios.

6 Discussion

Stratified Sampling Trade-offs: Biasing learning toward long-tail and edge cases could in principle sacrifice mainstream performance. However, we observe improvements across the board through two mechanisms. First, edge-case sampling reduces label noise in the negative class by filtering contamination in behavioral overlap regions where fraud and legitimate activity are difficult to distinguish. Second, floor constraints correct under-representation without over-representation, mainstream behaviors still dominate by volume, but edge cases are no longer systematically absent. This ensures the model learns both common and rare patterns.

Limitations: Concept and temporal drift remain ongoing challenges inherent to the fraud detection domain, requiring regular model updates. SAGE is not immune to these shifts despite reducing false positives on edge cases. The approach requires sufficient labeled fraud samples to fit the gates, and threshold tuning demands careful validation to balance contamination and yield.

Future Directions: Promising extensions include incorporating temporal graph structures to capture evolving fraud networks, applying contrastive learning to enhance edge case representation, and exploring additional gates (Isolation Forest, domain-specific heuristics) within the modular ensemble.

Acknowledgments

We thank the Amazon Music Product, Engineering, and Science teams for their support and collaboration on this work. We are especially grateful to the Operations team for their domain expertise, timely feedback, and detailed manual investigations, which were instrumental in shaping the labeling strategy and validating the model outputs. This project was made possible by the collective cross-functional effort and shared commitment to safeguarding the integrity of the Amazon Music streaming ecosystem.

References

[1] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2024. Self-Training: A Survey. arXiv:2202.12040 [cs.LG] https://arxiv.org/abs/2202.12040

[2] Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning* 109, 4 (2020), 719–760.

[3] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *ACM SIGMOD Record*, Vol. 29. 93–104.

[4] Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 380–388.

[5] Guangxin Chen, Fangqing Ye, Zuoyong Tian, Xuemin Zhu, and Qingming Huang. 2021. Positive-Unlabeled Learning from Imbalanced Data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*. 2995–3001. https://doi.org/10.24963/ijcai.2021/412

[6] CNM. 2023. Streaming fraud accounts for at least 1-3% of plays on services like Spotify and Deezer in France, shows investigation. https://www.musicbusinessworldwide.com/streaming-fraud-accounts-for-

at-least-1-3-of-plays-on-services-like-spotify-and-deezer-in-france-shows-investigation/. Accessed: 2023.

[7] Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications* 41, 10 (2014), 4915–4928.

[8] Thomas G Dietterich. 2000. Ensemble methods in machine learning. *Multiple classifier systems* (2000), 1–15.

[9] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 213–220.

[10] Soheil Esmaeilzadeh, Negin Salajegheh, Amir Ziai, and Jeff Boote. 2022. Abuse and Fraud Detection in Streaming Services Using Heuristic-Aware Machine Learning. (2022). arXiv:2203.02124 [cs.LG]

[11] Jonas Herskind Sejr, Thorbjørn Christiansen, Nicolai Dvinge, Dan Hougesen, Peter Schneider-Kamp, and Arthur Zimek. 2021. Outlier Detection with Explanations on Music Streaming Data: A Case Study with Danmark Music Group Ltd. *Applied Sciences* 11, 5 (2021). https://doi.org/10.3390/app11052270

[12] IFPI. 2025. Global Music Report 2025: Amidst Highly Competitive Market, Global Recorded Music Revenues Grew 4.8% in 2024. https://www.ifpi.org/ifpi-amidst-highly-competitive-market-global-recorded-music-revenues-grew-4-8-in-2024/. Accessed: 2025.

[13] Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (1998), 604–613.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, Vol. 30. 3146–3154.

[15] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[16] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems*. 1675–1685.

[17] Olivier Ledoit and Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis* 88, 2 (2004), 365–411.

[18] Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. 2002. Partially supervised classification of text documents. In *ICML*, Vol. 2. 387–394.

[19] Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. *National Institute of Science of India* (1936).

[20] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*. 141–150.

[21] Anand Muralidhar, Sharad Chitlangia, Rajat Agarwal, and Muneeb Ahmed. 2023. Real-time detection of robotic traffic in online advertising *(AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1775, 9 pages. https://doi.org/10.1609/aaai.v37i13.26844

[22] Music Business Worldwide. 2024. Streaming fraud costs the global music industry $2bn a year, according to Beatdapp. https://www.musicbusinessworldwide.com/streaming-fraud-costs-the-global-music-industry-2bn-a-year-according-to-beatdapp-now-its-partnering-with-beatport-to-combat-the-trend/. Accessed: 2024.

[23] Music In Africa. 2024. MLC and Beatdapp join forces to combat streaming fraud. https://www.musicinafrica.net/magazine/mlc-and-beatdapp-join-forces-combat-streaming-fraud. Accessed: 2024.

[24] Eric WT Ngai, Yong Hu, Yiu Hing Wong, Yijun Chen, and Xin Sun. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems* 50, 3 (2011), 559–569.

[25] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, Vol. 29. 427–438.

[26] RIAA. 2024. 2023 Year-End Revenue Statistics. https://www.riaa.com/wp-content/uploads/2024/03/2023-Year-End-Revenue-Statistics.pdf. Accessed: 2024.

[27] Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.

[28] U.S. Department of Justice. 2024. North Carolina Musician Charged in Music Streaming Fraud Aided by Artificial Intelligence. https://www.justice.gov/usao-sdny/pr/north-carolina-musician-charged-music-streaming-fraud-aided-artificial-intelligence. Accessed: 2024.

[29] David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Cambridge, Massachusetts, USA, 189–196. https://doi.org/10.3115/981658.981684

[30] Show-Jane Yen and Yue-Shi Lee. 2009. Cluster-based under-sampling approaches for imbalanced data distributions. In *Expert Systems with Applications*, Vol. 36. 5718–5727.