Adaptive Coreset Selection via Uncertainty-Density for Efficient Spam Detection (Student Abstract)

Aisha Hassan, Tushar Shinde

Indian Institute of Technology Madras, Zanzibar, Tanzania
zda24b009@iitmz.ac.in, shinde@iitmz.ac.in

Abstract

Efficient spam detection in resource-constrained environments remains challenging due to class imbalance, noisy text, and the computational demands of large Transformer models. We introduce a novel coreset selection framework based on a unified Entropy, Class-Balanced Uncertainty-Density Ranking (CBUDR) scheme. Our method prioritizes highly informative and uncertain samples while ensuring diversity and class balance within the selected subset. The framework flexibly supports multiple selection strategies, including Top-K, Bottom-K, and adaptive class-wise schemes, enabling robust performance even when training on as little as 5% of the dataset. Extensive experiments on benchmark datasets (UCI SMS, UTKML Twitter, LingSpam) show that our ranking scheme achieves competitive accuracy, precision, and recall while significantly reducing computational cost. These results demonstrate that carefully designed coreset strategies can surpass full-data performance in both balanced and imbalanced settings, highlighting the potential for deployment on low-power devices and mobile platforms.

Introduction

Spam detection is critical for user trust and service provider costs (Abdulhamid et al. 2017; Liu, Lu, and Nayak 2021; Al Saidat, Yerima, and Shaalan 2024). While Transformer-based models (e.g., BERT, RoBERTa) achieve state-of-the-art text classification, their reliance on large datasets makes training expensive and statistically inefficient (Devlin et al. 2019; Pal et al. 2025; Zhang et al. 2025). Our work specifically focuses on spam detection under limited labeled data (Oyeyemi and Ojo 2024; Xia and Chen 2020), rather than general coreset methods (Xia et al. 2023; Shinde and Madabhushi 2025; Shinde 2025; Shinde et al. 2025; Shinde and Sharma 2025).

Coreset selection aims to construct a smaller, representative subset of data that preserves model performance (Xia et al. 2023). Prior strategies include random sampling, uncertainty-driven (entropy/margin), and diversity-based clustering (Guo, Zhao, and Bai 2022), but these methods rarely optimize uncertainty and diversity jointly under class imbalance.

Contributions. We propose a multi-objective coreset selection framework that integrates statistical uncertainty and geometric representativeness with class-balancing. Our contributions are:

- We formulate coreset selection as a convex combination of class-normalized uncertainty and density-based representativeness.
- We introduce CBUDR, ensuring fairness across classes emphasizing informative, under-represented samples.
- We propose adaptive class-wise selection strategies to handle severe class imbalance.
- Experiments demonstrate that our method achieves competitive or superior performance while reducing training data by up to 95%.

Method

Let each sample x_i have embedding e_i and predictive probabilities $\mathbf{p}(x_i) = [p_1, \ldots, p_C]$. We define **Entropy-based Uncertainty.**

$$U(x_i) = -\sum_{c=1}^{C} p_c \log p_c, \tag{1}$$

which estimates the expected information gain from labeling x_i. High entropy identifies samples with maximal label uncertainty, crucial for exploration.

Class-Balanced Uncertainty-Density Ranking (CBUDR). To mitigate class imbalance, we normalize entropy within each class:

$$U_c(x_i) = \frac{U(x_i)}{\max_{x \in C_c} U(x)}. \tag{2}$$

Density-based representativeness ensures coverage of under-represented regions:

$$D(x_i) = 1 - \frac{1}{|N_i|} \sum_{x_j \in N_i} \text{sim}(e_i, e_j), \tag{3}$$

where N_i are k-nearest neighbors and sim is cosine similarity. The joint CBUDR score is:

$$\text{CBUDR}(x_i) = \alpha\, U_c(x_i) + \beta\, D(x_i), \quad \alpha + \beta = 1. \tag{4}$$

Dataset	Coreset Strategy	Ranking Method	5%				10%				25%			
			Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)
UtkMl Twitter	Random		94.44	93.98	**100.00**	88.64	94.44	94.12	97.56	90.91	95.55	95.41	98.11	92.86
	Class-wise Top-K	Entropy	63.33	67.33	59.65	77.27	83.33	81.01	90.14	73.56	90.42	90.02	91.08	88.99
		CBUDR	73.33	68.42	81.25	59.09	89.44	89.14	88.64	89.66	88.20	88.40	84.52	92.66
		Entropy+CBUDR	78.89	76.54	83.78	70.45	82.22	78.67	93.65	67.82	89.76	89.55	88.74	90.37
	Class-wise Bottom-K	Entropy	98.89	*98.85*	**100.00**	97.73	*98.33*	98.27	*98.84*	97.70	98.44	98.38	99.07	97.71
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**99.44**	**99.42**	**100.00**	98.85	**99.33**	99.31	**100.00**	98.62
		Entropy+CBUDR	98.89	98.88	97.78	**100.00**	*98.33*	98.25	**100.00**	96.55	*98.66*	98.61	99.53	97.71
	Baseline	All (100%)									96.49	96.41	95.92	96.91
UCI	Random	None	**100.00**	**100.00**	**100.00**	**100.00**	97.62	91.67	84.62	**100.00**	98.09	93.10	*90.00*	96.43
	Class-wise Top-K	Entropy	*90.48*	*60.00*	*60.00*	*60.00*	90.48	63.64	*63.64*	63.64	*99.04*	*96.30*	**100.00**	*92.86*
		CBUDR	*90.48*	33.33	**100.00**	20.00	*91.67*	*74.07*	62.50	90.91	97.61	90.91	92.59	89.29
		Combined	*90.48*	33.33	**100.00**	20.00	90.48	69.23	60.00	*81.82*	97.13	88.00	**100.00**	78.57
	Class-wise Bottom-K	Entropy	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.52	98.18	**100.00**	96.43
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
		Combined	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.52	98.18	**100.00**	96.43
	Baseline	All (100%)									99.52	98.18	**100.00**	96.43
LingSpam	Random	None	90.91	50.00	**100.00**	33.33	88.64	*70.59*	*60.00*	85.71	*99.08*	*97.30*	*94.74*	**100.00**
	Class-wise Top-K	Entropy	*86.36*	*0.00*	0.00	*0.00*	*81.82*	55.56	45.45	*71.43*	87.16	61.11	61.11	61.11
		CBUDR	90.91	50.00	**100.00**	33.33	79.55	40.00	37.50	42.86	93.58	78.79	86.67	*72.22*
		Combined	81.82	*0.00*	0.00	*0.00*	77.27	37.50	33.33	42.86	95.41	83.87	**100.00**	*72.22*
	Class-wise Bottom-K	Entropy	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	**93.33**	87.50	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
		CBUDR	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	92.31	**100.00**	85.71	**100.00**	**100.00**	**100.00**	**100.00**
		Combined	**100.00**	**100.00**	**100.00**	**100.00**	**97.73**	92.31	**100.00**	85.71	**100.00**	**100.00**	**100.00**	**100.00**
	Baseline	All (100%)									99.54	98.61	98.61	98.61

Table 1: Performance of Different Coreset Selection Strategies and Ranking Methods on different datasets.

Entropy + CBUDR Combination. To further enhance the coreset selection, we combine entropy and CBUDR scores. The combined uncertainty score $U'(x_i)$ is calculated by weighting entropy score and CBUDR score as follows:

$$U'(x_i) = \lambda U(x_i) + (1-\lambda)\,\text{CBUDR}(x_i), \quad \lambda \in [0,1], \quad (5)$$

balancing global exploration and class-balanced coverage. Top-$K\%$ of $U'(x_i)$ forms the coreset, approximating an information-theoretically optimal distribution. Time complexity is $O(nkd)$ for n samples with d-dimensional embeddings. Scalable implementations using approximate nearest neighbors are feasible for large datasets (Hassan and Shinde 2025b,c,a).

Experimental Setup

Datasets. We benchmark on three spam datasets with varying noise and imbalance: UTKML Twitter (11,968 tweets, balanced), UCI SMS (5,572 messages, 13% spam), LingSpam (2,893 emails, 16% spam). Messages are tokenized, lowercased, and embedded via SBERT.

Coreset Selection and Evaluation. Ranking methods: Entropy, CBUDR, Entropy+CBUDR. Selection strategies: Random, Top-K, Bottom-K, Classwise Top/Bottom-K. Metrics: Accuracy, Precision, Recall, F1-score, averaged over three seeds.

Results and Discussion

We evaluate different coreset selection strategies: Top-K, Bottom-K, and Class-wise, combined with Entropy, CBUDR, and their combination on UTKML Twitter Spam, UCI SMS Spam, and Ling-Spam datasets. Coreset sizes of 5%, 10%, and 25% are considered to identify data-efficient strategies that maintain or exceed full-dataset performance, addressing class imbalance.

UTKML: Bottom-K consistently outperforms Top-K, e.g., Bottom-K (Entropy) achieves F1=98.38% at 25%, exceeding the full-data baseline of 96.41%. Entropy-based Top-K favors recall but reduces precision, while CBUDR provides balanced trade-offs. Combined entropy+CBUDR improves robustness (F1=98.61%), and Class-wise stabilizes recall.

UCI SMS: Entropy Top-K underperforms (F1=60.00% at 5%) due to majority-class bias. CBUDR Top-K improves balance (Precision=62.50%, Recall=90.91% at 10%). Bottom-K dominates across coreset sizes, achieving near-perfect F1 and perfect precision/recall at 25%. Class-wise strategies stabilize minority-class recall but slightly lag behind Bottom-K.

Ling-Spam: Entropy Top-K fails at small coresets (F1=0% at 5%), while CBUDR and combined ranking improve performance (F1=83.87% at 25%). Bottom-K maintains near-perfect metrics across all sizes, and Class-wise Bottom-K preserves minority-class inclusion.

Bottom-K, especially with CBUDR, consistently outperforms other strategies and often exceeds full-dataset training. Entropy alone is insufficient under imbalance. Combining entropy and CBUDR enhances robustness at larger coresets, while Class-wise selection ensures minority-class representation. Efficient coreset design allows up to 95% dataset reduction without performance loss.

Conclusion

We proposed a coreset selection method for efficient spam detection using the Entropy + Density Uncertainty Ranking (EDUR) framework, combining predictive uncertainty and sample representativeness. We explicitly address spam detection and demonstrate method scalability. Experiments on benchmark spam datasets show that EDUR reduces training data by up to 75% without sacrificing accuracy, precision, or recall. Bottom-K strategies consistently achieve near-perfect F1-scores, proving effective in both balanced and highly imbalanced scenarios. This makes EDUR suitable for resource-constrained environments such as mobile or low-power systems. Integrating active learning techniques, exploring hybrid uncertainty measures, and applying EDUR to domains like fraud detection, phishing, or misinformation filtering can further enhance selection efficiency and robustness.

References

Abdulhamid, S. M.; Abd Latiff, M. S.; Chiroma, H.; Osho, O.; Abdul-Salaam, G.; Abubakar, A. I.; and Herawan, T. 2017. A review on mobile SMS spam filtering techniques. *IEEE Access*, 5: 15650–15666.

Al Saidat, M. R.; Yerima, S. Y.; and Shaalan, K. 2024. Advancements of SMS spam detection: A comprehensive survey of NLP and ML techniques. *Procedia Computer Science*, 244: 248–259.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Guo, C.; Zhao, B.; and Bai, Y. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, 181–195. Springer.

Hassan, A. H.; and Shinde, T. 2025a. Efficient Spam Detection with Sentence-BERT using Adaptive Uncertainty-Diversity Ranking Coresets. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

Hassan, A. H.; and Shinde, T. 2025b. Optimized Statistical Ranking is All You Need for Robust Coreset Selection in Efficient Transformer-Based Spam Detection. In *OPT 2025: Optimization for Machine Learning*.

Hassan, A. H.; and Shinde, T. 2025c. Uncertainty-Diversity Ranking Coreset Selection for Efficient Spam Detection. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*.

Liu, X.; Lu, H.; and Nayak, A. 2021. A spam transformer model for SMS spam detection. *IEEE Access*, 9: 80253–80263.

Oyeyemi, D. A.; and Ojo, A. K. 2024. SMS Spam Detection and Classification to Combat Abuse in Telephone Networks Using Natural Language Processing. *arXiv preprint arXiv:2406.06578*.

Pal, A. A.; Mondal, S.; Kumar, C. A.; and Kumar, C. J. 2025. A Transformer-Based Approach for Fake News and Spam Detection in Social Media Using RoBERTa. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, 1256–1263. IEEE.

Shinde, T. 2025. High-Performance Lightweight Vision Models for Land Cover Classification with Coresets and Compression. In *TerraBytes-ICML 2025 workshop*.

Shinde, T.; and Madabhushi, M. 2025. Data-Efficient and Robust Coreset Selection via Sparse Adversarial Perturbations. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.

Shinde, T.; and Sharma, A. K. 2025. Scalable and Efficient Multi-Weather Classification for Autonomous Driving with Coresets, Pruning, and Resolution Scaling. In *ICLR 2025 Workshop on Machine Learning Multiscale Processes*.

Shinde, T.; Sharma, A. K.; Bhardwaj, S.; and Vuai, A. S. 2025. Navigating Coreset Selection and Model Compression for Efficient Maritime Image Classification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 1608–1616.

Xia, T.; and Chen, X. 2020. A discrete hidden Markov model for SMS spam detection. *Applied Sciences*, 10(14): 5011.

Xia, X.; Liu, J.; Zhang, S.; Wu, Q.; Wei, H.; and Liu, T. 2023. Refined coreset selection: Towards minimal coreset size under model performance constraints. *arXiv preprint arXiv:2311.08675*.

Zhang, H.; Liu, Y.; Qiu, Y.; Liu, H.; Pei, Z.; Wang, J.; and Long, M. 2025. Timesbert: A bert-style foundation model for time series understanding. *arXiv preprint arXiv:2502.21245*.