Uncertainty-Diversity Ranking Coreset Selection for Efficient Spam Detection

Aisha Hamad Hassan, Tushar Shinde
MIDAS (Multimedia Intelligence, Data Analysis and compreSsion) Lab
Indian Institute of Technology Madras, Zanzibar, Tanzania
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 Uncertainty-Diversity Ranking (UDR), which explicitly combines predictive uncertainty with representativeness to prioritize highly informative samples while ensuring diversity and class balance. Our method supports multiple coreset strategies, including Top-K, Bottom-K, and adaptive class-wise selection, enabling robust performance even with a fraction of the training data. Extensive experiments on benchmark datasets, including UCI SMS, UTKML Twitter, and Ling-Spam, show that UDR maintains or improves accuracy, precision, and recall while reducing training data by up to 95%, significantly lowering computational cost. These results demonstrate the potential of UDR in resource-limited settings.

1 Introduction and Related Work

Spam detection, particularly in mobile SMS filtering, has attracted significant attention due to its impact on user trust and the financial burden it imposes on service providers Abdulhamid et al. [2017], Liu et al. [2021], Al Saidat et al. [2024]. Despite advances in natural language processing (NLP), spam detection remains challenging due to severe class imbalance, noisy and unstructured text distributions, and the scarcity of large-scale labeled datasets that preserve real-world statistics Oyeyemi and Ojo [2024], Xia and Chen [2020]. These challenges exemplify broader issues in machine learning where efficient training must balance predictive uncertainty with representative coverage under limited data, a setting directly relevant to uncertainty-aware operational decision-making in resource-constrained environments. Transformer-based models such as BERT and RoBERTa achieve state-of-the-art text classification performance Devlin et al. [2019], Pal et al. [2025], Zhang et al. [2025]. However, their reliance on large datasets makes training both computationally expensive and statistically inefficient Abdulhamid et al. [2017]. This motivates the study of data-efficient optimization strategies that reduce the dataset size while preserving generalization and robust decision-making.

Coreset selection addresses this problem by constructing a smaller, informative subset of the training data that approximates the performance of the full dataset Xia et al. [2023], Shinde and Madabhushi, Shinde, Shinde et al. [2025], Shinde and Sharma. Existing strategies include random sampling, uncertainty-driven approaches (e.g., entropy or margin-based selection), and diversity-based clustering techniques Guo et al. [2022], Chai et al. [2023]. While effective individually, these methods struggle to jointly optimize for uncertainty (capturing statistically ambiguous samples) and diversity (ensuring coverage of the underlying distribution), especially under class imbalance. We propose a novel framework that balances these two objectives by prioritizing samples that are both uncertain and representative, ensuring that the selected subset captures the key decision-critical information of the full dataset.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Our Contributions. We propose an optimization-driven coreset selection framework that unifies entropy-based uncertainty with density-aware representativeness:

- We formalize coreset selection as a unified approach that balances statistical uncertainty and geometric coverage, bridging uncertainty sampling with distributional representativeness.
- We introduce Class-Balanced Uncertainty-Density Ranking (CBUDR), which normalizes uncertainty within each class and incorporates density to improve stability under imbalance.
- We extend the framework with adaptive class-wise selection schemes, enabling targeted sampling for rare classes and efficient data-driven decision-making.

2 Method

We propose a novel coreset selection framework that jointly integrates statistical uncertainty with geometric representativeness under a class-balanced constraint. Specifically, we design two ranking functions, Entropy-based Uncertainty and Class-Balanced Uncertainty-Density Ranking (CBUDR), and unify them through a convex combination. This approach can be interpreted as a multi-objective strategy balancing exploration (selecting uncertain samples) and coverage (selecting representative samples), which is crucial for uncertainty-aware decision-making in resource-limited settings.

Entropy-based Ranking. Each input sample x_i is mapped to an embedding e_i via a pretrained encoder (e.g., SBERT). A proxy classifier produces predictive probabilities $\mathbf{p}(x_i) = [p_1, \ldots, p_C]$ over C classes. The uncertainty of x_i is quantified via Shannon entropy: $U(x_i) = -\sum_{c=1}^{C} p_c \log p_c$. Entropy measures the expected information gain from querying the label of x_i. High entropy identifies samples that are ambiguous to the model, making them most informative for reducing predictive uncertainty.

Class-Balanced Uncertainty-Density Ranking (CBUDR). Entropy alone may over-prioritize minority or noisy regions. To mitigate this, we normalize entropy within each class. For sample $x_i \in C_c$: $U_c(x_i) = \frac{U(x_i)}{\max_{x \in C_c} U(x)}$. Normalization ensures that each class contributes comparably to the selection, reducing bias from class imbalance. To capture geometric coverage, we compute a density score: $D(x_i) = 1 - \frac{1}{|N_i|} \sum_{x_j \in N_i} \text{sim}(e_i, e_j)$, where N_i denotes the k-nearest neighbors of x_i in embedding space and sim is cosine similarity. A higher density score indicates that x_i lies in a sparsely populated region, ensuring the coreset spans the input manifold and captures diverse patterns. For joint ranking, these components are combined as a convex score:

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

This combination provides a controllable trade-off between exploration (uncertainty) and coverage (representativeness), allowing systematic navigation of the uncertainty-diversity frontier.

Entropy + CBUDR Combination. While CBUDR balances class-wise uncertainty and representativeness, global entropy captures informativeness across all classes. We unify these perspectives via:

$$U'(x_i) = \lambda \cdot U(x_i) + (1 - \lambda) \cdot \text{CBUDR}(x_i), \quad \lambda \in [0, 1]. \tag{2}$$

The parameter λ controls the relative emphasis on global exploration versus class-aware coverage. Ranking samples by $U'(x_i)$ and selecting the top-$K\%$ yields a coreset that both reduces predictive entropy and ensures geometric diversity, improving convergence speed and generalization under limited data. This helps by balancing statistical efficiency and robust decision-making under uncertainty.

3 Experimental Setup

We evaluate our coreset selection strategy on three benchmark spam detection datasets: **UTKML Twitter Spam** (11,968 timestamped tweets), **UCI SMS Spam Collection** (5,572 sequential messages), and **Ling-Spam** (2,893 time-ordered emails). The **UTKML Twitter Spam** dataset (48.6% spam) is approximately balanced, while the **UCI SMS Spam Collection** (13.4% spam) and **Ling-Spam** (16.6% spam) are imbalanced, providing a testbed for uncertainty-aware and class-sensitive selection strategies. *Preprocessing:* Messages are tokenized, lowercased, and stripped of punctuation and stopwords. We additionally handle social media-specific tokens (e.g., URLs, hashtags, mentions)

to preserve semantic meaning. Each message is embedded using Sentence-BERT (SBERT) Reimers and Gurevych [2019], producing semantically meaningful vectors that are essential for the density-based selection in CBUDR. Without such embeddings, similarity measures would poorly reflect text semantics, undermining representativeness.

Coreset Selection. We compare Random, Top-K, Bottom-K, and Class-wise Top-K/Bottom-K strategies using entropy, CBUDR, and their combination. Coresets are selected to optimize training efficiency and transferability. Entropy and CBUDR scores are computed as described in Section 2, and adaptive strategies dynamically adjust λ based on validation performance, improving model generalization across sequential tasks.

Model Training. All experiments are conducted on the Kaggle platform using an NVIDIA Tesla P100 GPU. Each coreset is used to fine-tune a pretrained BERT model. Data is split 75% train, 15% validation, and 15% test. Training uses the Adam optimizer (learning rate 2e-5) with a batch size of 32. Fine-tuning on coresets rather than the full dataset simulates resource-constrained scenarios and evaluates how much performance is retained under aggressive data reduction. As an upper bound, we also fine-tune on the full dataset.

Evaluation Metrics. Performance is measured using accuracy, F1-score, precision, and recall. For imbalanced datasets, recall is emphasized to reflect the practical importance of correctly detecting minority (spam) classes, aligning with uncertainty-aware decision-making principles.

4 Results and Discussion

We evaluate the effect of different coreset selection strategies across three benchmark datasets. We compare Top-K, Bottom-K, and Class-wise selection combined with Entropy, CBUDR, and their combination at coreset sizes of 5%, 10%, and 25%. *Coreset Scoring Mechanism:* Each sample is scored based on predictive uncertainty and representativeness in embedding space. Top-K prioritizes the most informative samples, Bottom-K emphasizes the least ambiguous samples, and Class-wise selection ensures balanced coverage across classes.

UTKML Twitter Spam (Balanced Dataset). Table 1 summarizes results for the UTKML dataset. Bottom-K consistently outperforms Top-K, suggesting that emphasizing "easy" low-entropy samples improves generalization even in balanced settings. CBUDR enhances this effect by maintaining geometric coverage, and the combination of entropy and CBUDR further stabilizes performance. Class-wise selection ensures fair representation of spam and ham, slightly reducing F1 compared to Bottom-K but stabilizing minority-class recall. Overall, well-designed coresets can surpass full-data performance even with 95% data reduction.

UCI SMS Spam and Ling-Spam (Imbalanced Datasets). For UCI SMS Spam (13% spam) and Ling-Spam (highly imbalanced), Entropy-based Top-K struggles under severe class imbalance, over-selecting majority-class samples and degrading minority-class performance. CBUDR improves outcomes by prioritizing representative dense samples, achieving high recall without sacrificing precision. Bottom-K consistently dominates across coreset sizes, preserving clear decision boundaries and near-perfect F1-scores, while class-wise selection ensures minority-class inclusion and stability. These results highlight that principled coreset design is essential for robust, data-efficient learning in resource-constrained, uncertainty-aware settings.

Table 1: Performance of Different Coreset Selection Strategies and Ranking Methods on UtkMl Twitter Spam Dataset.

Coreset Strategy	Ranking Method	5%				10%				25%			
		Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)
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
Top-K	Entropy	71.11	74.51	66.67	84.44	87.22	86.55	93.67	80.43	88.86	88.89	86.58	91.32
Top-K	CBUDR	67.78	60.27	70.97	52.38	85.56	85.56	85.56	85.56	89.31	88.84	90.95	86.82
Top-K	Entropy+CBUDR	78.89	73.24	*96.30*	59.09	83.33	84.69	80.58	89.25	91.76	91.90	88.98	95.02
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
Class-wise Top-K	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
Class-wise Top-K	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
Bottom-K	Entropy	*97.78*	98.11	*96.30*	**100.00**	*98.33*	*98.52*	99.01	98.04	98.22	98.33	97.51	99.16
Bottom-K	CBUDR	98.89	98.41	**100.00**	96.88	98.89	98.78	98.78	*98.78*	98.89	*98.92*	99.14	*98.71*
Bottom-K	Entropy+CBUDR	98.89	98.80	**100.00**	*97.62*	97.78	97.85	96.81	**98.91**	**99.33**	**99.36**	*99.15*	**99.57**
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
Class-wise Bottom-K	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
Class-wise Bottom-K	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

Table 2: Performance of Different Coreset Selection Strategies and Ranking Methods on UCI Dataset.

Coreset Strategy	Ranking Method	5%				10%				25%			
		Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)
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
Top-K	Entropy	92.86	72.73	66.67	80.00	90.48	63.64	*63.64*	63.64	*99.04*	*96.30*	**100.00**	*92.86*
Top-K	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
Top-K	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 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*
Class-wise Top-K	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
Class-wise Top-K	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
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
Bottom-K	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**
Bottom-K	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
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
Class-wise Bottom-K	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**
Class-wise Bottom-K	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

Table 3: Performance of Different Coreset Selection Strategies and Ranking Methods on LingSpam Dataset.

Coreset Strategy	Ranking Method	5%				10%				25%			
		Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Acc (%)	F1 (%)	Prec (%)	Rec (%)
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**
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
Top-K	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*
Top-K	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 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
Class-wise Top-K	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*
Class-wise Top-K	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*
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**
Bottom-K	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**
Bottom-K	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**
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**
Class-wise Bottom-K	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**
Class-wise Bottom-K	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

Discussion. Across all datasets, several trends emerge. Bottom-K strategies, particularly when combined with CBUDR, consistently outperform other approaches and frequently surpass full-dataset baselines, demonstrating that selecting "easy" yet representative samples can improve generalization under limited data. Entropy alone is insufficient under imbalance, over-selecting ambiguous majority-class instances. CBUDR contributes complementary benefits by emphasizing representativeness and density. The hybrid entropy+CBUDR scoring further enhances robustness, especially at larger coreset sizes. Class-wise selection guarantees minority-class inclusion, critical under severe imbalance. Overall, these findings demonstrate that careful coreset design enables dataset reductions of up to 95% while maintaining or improving predictive performance, providing a practical and principled approach to uncertainty-aware, resource-efficient learning in challenging real-world scenarios.

5 Conclusion

We presented CBUDR, a novel coreset selection strategy combining predictive uncertainty (entropy) with representativeness (density) for efficient spam detection. CBUDR enables models to focus on decision-critical samples while reducing redundant or noisy data. Experiments on multiple datasets show up to 95% training reduction with maintained or improved F1-scores. Bottom-K strategies excel by emphasizing "easy" yet representative samples, balancing exploration and coverage across balanced and imbalanced data. CBUDR is suitable for resource-constrained settings, preserving predictive reliability while reducing computational and memory demands. **Future Directions.** We plan to explore adaptive, dynamic coreset construction via active learning and hybrid uncertainty measures, and to extend CBUDR to fraud, phishing, misinformation, and adversarial detection, validating robustness under real-world shifts.

References

Shafi'I Muhammad Abdulhamid, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I Abubakar, and Tutut Herawan. A review on mobile sms spam filtering techniques. *IEEE Access*, 5:15650–15666, 2017.

Mohammed Rasol Al Saidat, Suleiman Y Yerima, and Khaled Shaalan. Advancements of sms spam detection: A comprehensive survey of nlp and ml techniques. *Procedia Computer Science*, 244:248–259, 2024.

Chengliang Chai, Jiayi Wang, Nan Tang, Ye Yuan, Jiabin Liu, Yuhao Deng, and Guoren Wang. Efficient coreset selection with cluster-based methods. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 167–178, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022.

Xiaoxu Liu, Haoye Lu, and Amiya Nayak. A spam transformer model for sms spam detection. *IEEE Access*, 9: 80253–80263, 2021.

Dare Azeez Oyeyemi and Adebola K Ojo. Sms spam detection and classification to combat abuse in telephone networks using natural language processing. *arXiv preprint arXiv:2406.06578*, 2024.

Ankit Abhijit Pal, Sudin Mondal, C Ashok Kumar, and C Jothi Kumar. 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)*, pages 1256–1263. IEEE, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Tushar Shinde. High-performance lightweight vision models for land cover classification with coresets and compression. In *TerraBytes-ICML 2025 workshop*.

Tushar Shinde and Manasa Madabhushi. Data-efficient and robust coreset selection via sparse adversarial perturbations. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.

Tushar Shinde and Avinash Kumar Sharma. Scalable and efficient multi-weather classification for autonomous driving with coresets, pruning, and resolution scaling. In *ICLR 2025 Workshop on Machine Learning Multiscale Processes*.

Tushar Shinde, Avinash Kumar Sharma, Shivam Bhardwaj, and Ahmed Silima Vuai. Navigating coreset selection and model compression for efficient maritime image classification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1608–1616, 2025.

Tian Xia and Xuemin Chen. A discrete hidden markov model for sms spam detection. *Applied Sciences*, 10(14): 5011, 2020.

Xiaobo Xia, Jiale Liu, Shaokun Zhang, Qingyun Wu, Hongxin Wei, and Tongliang Liu. Refined coreset selection: Towards minimal coreset size under model performance constraints. *arXiv preprint arXiv:2311.08675*, 2023.

Haoran Zhang, Yong Liu, Yunzhong Qiu, Haixuan Liu, Zhongyi Pei, Jianmin Wang, and Mingsheng Long. Timesbert: A bert-style foundation model for time series understanding. *arXiv preprint arXiv:2502.21245*, 2025.