Active Learning for Imbalanced Biomedical Text Classification: Reducing Annotation Costs in Small Sample Settings

22AIE213- Machine Learning

TEAM 12

CH.SC.U4AIE23009 – CH.Naynesh Reddy
CH.SC.U4AIE23037 -- N.Bhargav
CH.SC.U4AIE23040 – P.MD.Kalesha

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

Literature Review

S.NO	Title	Citation Number	Methodology	Pros	Cons	Research gap
1	Active Learning for Biomedical Article Classification with Bag of Words and *Fast Text* Embeddings	https://doi.org/10.3390/app14177945	Active learning with BoW and FastText embeddings selects uncertain samples for labeling, improving classification efficiency while reducing annotation effort.	• Cuts labeling effort by 50% while keeping accuracy high. • Works well for biomedical text classification.	• Tested less on highly imbalanced datasets. • Depends on BoW & FastText quality.	• Needs scalability testing on larger datasets. • Impact on high-dimensional data remains unexplored.- Lacks validation.
2	Class Balancing for Efficient Active Learning in Imbalanced Datasets	https://aclanthology.org/2024.law-1.8/	Introduces a tune-free weighting method that integrates class balancing into active learning algorithms for improved sample selection.	• Eliminates manual tuning of weighting functions. • Improves performance without requiring external balancing techniques.	• May not generalize well to highly dynamic datasets. • Requires additional computational overhead.	• Needs further exploration in real-world classification tasks. • Impact of noisy labels on the method remains untested.
3	Algorithm Selection for Deep Active Learning with Imbalanced Datasets	https://proceedings.neurips.cc/paper_files/paper/2023/file/1e77af93008ee6cd248a31723ce357d8-Paper-Conference.pdf	Proposes TAILOR, a meta-algorithm that adaptively selects active learning strategies to balance sample acquisition.	• Achieves more efficient class-balanced sampling. • Reduces labeling costs in deep learning applications.	• Computationally expensive for large datasets. • Relies on predefined heuristics for model selection.	• Needs real-world benchmarking across multiple domains. • Requires further validation on long-tailed distributions.
4	BASIL: Balanced Active Semi-supervised Learning for Class Imbalanced Datasets	https://doi.org/10.48550/arXiv.2203.05651	Uses submodular mutual information functions to select a balanced dataset in an active learning loop.	• Improves label efficiency by integrating semi-supervised learning. • Reduces the need for manually labeled data.	• Sensitive to initial model performance. • Performance depends on dataset diversity.	• Needs testing on larger, more complex datasets. • Impact of model drift over multiple iterations is unexplored.
5	DIRECT: Deep Active Learning under Imbalance and Label Noise	https://doi.org/10.48550/arXiv.2312.09196	Introduces a threshold-based annotation method to minimize annotation costs while handling class imbalance and label noise.	• Reduces annotation budget significantly. • Enhances learning for rare classes.	• High computational cost. • Struggles with extremely noisy datasets.	• Requires improvements in handling highly dynamic data streams. • Needs further evaluation on real-world imbalanced datasets.

S.NO	Title	Citation Number	Methodology	Pros	Cons	Research gap
6	Active Learning for Imbalanced Data Under Cold Start	https://doi.org/10.1145/3490354.3494423	Proposes an Outlier-based Discriminative Active Learning (ODAL) approach with a three-stage AL policy to address extreme class imbalance in cold start scenarios.	• Effective in low-data environments. • Achieves faster model convergence than random sampling.	• Performance depends on outlier detection accuracy. • May struggle with rare class generalization.	• Lacks deep learning embeddings like BERT. • Needs adaptation to specific biomedical datasets.
7	Active Learning for Imbalanced Datasets	https://openaccess.thecvf.com/content_WACV_2020/papers/Aggarwal_Active_Learning_for_Imbalanced_Datasets_WACV_2020_paper.pdf	Modifies active learning acquisition functions using a pretrained deep model and integrates a balancing step for improved class imbalance handling.	• Reduces annotation costs while improving minority class recall. • Adaptive selection of underrepresented samples.	• Requires extensive hyperparameter tuning. • Limited generalizability to highly skewed datasets.	• Needs validation on real-world industry datasets. • Computational efficiency for large datasets is unexplored.
8	Cost-Aware Active Learning for Named Entity Recognition in Clinical Text	https://doi.org/10.1093/jamia/ocz102	A cost-aware active learning algorithm (Cost-CAUSE) balances annotation cost and sample informativeness, optimizing selection to reduce labeling effort in clinical text NER.	• Enhances sample selection efficiency. • Adaptable to different deep learning architectures.	• Limited to named entity recognition tasks. • Effectiveness depends on accurate cost modeling.	• Needs testing on broader biomedical classification tasks. • Requires adaptation to deep learning-based NER models.
9	Deep Active Learning Models for Imbalanced Image Classification	https://doi.org/10.1016/j.knosys.2022.109817	Introduces a Balanced Active Learning (BAL) method for improving minority class representation in image classification.	• Cuts annotation effort while maintaining model accuracy. • Applies active learning across multiple biomedical tasks.	• Requires large-scale training data. • Sensitive to hyperparameter tuning.	• Needs cross-domain testing. • Effectiveness in real-time applications remains uncertain.
10	Active Learning with Applications in Biomedical Document Annotation	10.32657/10356/71680	Explores various active learning methods to reduce manual annotation in biomedical document tasks like event extraction and named entity recognition.	• Enhances active learning by integrating class balancing. • Improves accuracy on imbalanced image datasets.	• Focuses on specific annotation tasks, limiting generalization. • Performance varies with dataset characteristics.	• Needs integration with deep learning-based NLP models. • Lacks evaluation on large-scale biomedical datasets.

S.NO	Title	Citation Number	Methodology	Pros	Cons	Research gap
11	A Generic Semi-Supervised and Active Learning Framework for Biomedical Text Classification	10.1109/EMBC48229.2022.9871846	Combines active learning with semi-supervised learning to minimize manual labeling while improving biomedical text classification performance.	• Reduces labeling effort by 10% without affecting accuracy. • Works across various biomedical text classification tasks.	• Limited evaluation on highly imbalanced datasets. • Performance depends on initial labeled data quality.	• Needs integration with transformer-based models like BERT. • Requires testing on diverse biomedical text corpora.
12	CLINICAL: Targeted Active Learning for Imbalanced Medical Image Classification	10.1007/978-3-031-16760-7_12	Uses submodular mutual information to select rare class samples, improving classification in imbalanced datasets.	• Selects key data points from rare classes. • Outperforms existing methods in long-tail cases.	• Limited to medical image classification. • Needs fine-tuning of acquisition functions.	• Requires adaptation for biomedical text data. • Needs testing on diverse clinical datasets.
13	An Active Learning Based Classification Strategy for the Minority Class Problem: Application to Histopathology Annotation	10.1186/1471-2105-12-424	Uses class-balanced active learning (CBAL) to improve cancer detection in histopathology images.	• Boosts classifier accuracy for cancer detection. • Predicts annotation needs for efficient training.	• Focused only on histopathology images. • Depends on accurate class balance estimation.	• Needs application to other biomedical fields. • Requires integration with deep learning models.
14	AnchorAL: Computationally Efficient Active Learning for Large and Imbalanced Datasets	10.18653/v1/2024.naacl-long.467	Uses class-specific "anchor" instances from labeled data to find similar unlabeled samples, forming small, focused subpools for efficient active learning in imbalanced datasets.	• Reduces computation by limiting search space. • Improves minority class representation.	• Effectiveness depends on anchor selection quality. • Limited testing on biomedical text classification.	• Needs validation on biomedical datasets. • Requires integration with deep learning models.
15	Class Balancing for Efficient Active Learning in Imbalanced Datasets	10.18653/v1/2024.law-1.8	Introduces a tune-free weighting technique to balance class representation in active learning, ensuring diverse and representative sample selection.	• Improves active learning on imbalanced datasets. • No manual tuning required.	• Performance depends on dataset distribution. • Needs more evaluation on extreme imbalances.	• Requires testing on large-scale biomedical datasets. • Needs adaptation for real-world annotation settings.

S.NO	Title	Citation Number	Methodology	Pros	Cons	Research gap
16	Active Learning for Imbalanced Civil Infrastructure Data	10.48550/arXiv.2210.10586	Replaces standard acquisition functions with an auxiliary binary discriminator to improve active learning in imbalanced datasets, ensuring better sample selection.	• Reduces bias in active learning selections. • Works well with large labeled datasets.	• Focused on civil infrastructure, limiting biomedical applications. • Requires adaptation for text and genomic data.	• Needs validation in biomedical datasets. • Requires integration with transformer-based models.
17	Boosted SVM with Active Learning Strategy for Imbalanced Data	https://doi.org/10.1109/JIOT.2023.3260722	Implemented a boosted SVM with active learning to optimize classification on imbalanced data by selecting informative samples.	• Enhances SVM training with active learning for imbalanced datasets. • Reduces redundant data points while improving classification accuracy.	• Performance depends on dataset characteristics. • Requires careful parameter tuning for best results.	• Lacks real-world applications across diverse domains. • Needs comparison with other imbalance-handling methods.
18	On the Importance of Adaptive Data Collection for Extremely Imbalanced Pairwise Tasks	10.18653/v1/2020.findings-emnlp.305.	Designed an adaptive sampling method to collect informative negative examples, improving pairwise classification accuracy.	• Improves negative sample selection for pairwise classification. • Increases model precision through active learning.	• May not generalize well to all pairwise tasks. • Risk of bias in negative example selection.	• Limited application beyond pairwise tasks. • Requires testing across different levels of class imbalance.
19	Class-Balanced Active Learning for Image Classification	10.1109/WACV51458.2022.00376.	Integrated class balancing into active learning for image classification, ensuring equal representation and improving robustness.	• Enhances active learning by integrating class balancing. • Improves accuracy on imbalanced image datasets.	• Requires more computational resources. • Challenging to apply beyond image classification.	• Needs validation on non-image datasets. • Requires comparisons with alternative active learning methods.
20	Multiple Sampling Schemes and Deep Learning Improve Active Learning Performance in Drug-Drug Interaction Information Retrieval Analysis from the	10.1186/s13326-023-00287-7	Combines deep learning with multiple sampling strategies (random negative, positive, similarity, and uncertainty sampling) to improve active learning in drug-drug interaction extraction.	• Enhances efficiency in biomedical literature retrieval. • Uses diverse sampling strategies for better model training.	• Requires fine-tuning for different biomedical tasks. • Performance varies across sampling strategies.	• Needs evaluation on real-world clinical datasets. • Requires integration with more advanced deep learning models.

Problem Identification

Issue Addressed:

1. Annotating large datasets for imbalanced classes is costly and time-consuming.
2. Many instances do not contribute significantly to model learning, leading to unnecessary annotation efforts.
3. Traditional data annotation strategies fail to leverage active learning effectively, resulting in inefficient dataset utilization.

Gaps/Challenges Identified in Literature:

1. Existing annotation techniques treat all samples equally, leading to wasted labeling efforts.
2. Standard machine learning models struggle with imbalanced data without extensive labeled datasets.
3. Active learning methods are underutilized, and their potential to optimize annotation costs remains unexplored in many domains.

Problem Statement

Clear Problem Description:

Annotating large datasets for imbalanced classes is costly and often unnecessary when active learning is underutilized.

Solution Approach:

Use uncertainty sampling and diversity-based active learning to maximize learning efficiency. Apply support vector machines or tree-based models for scalability.

Objective

1. **Efficient Label Utilization:** Develop an active learning pipeline that reduces the number of labeled samples by 40-50% while maintaining classification accuracy for imbalanced biomedical datasets.

2. **Optimized Sampling Strategy:** Implement a hybrid sampling approach combining uncertainty-based selection and diversity sampling to maximize learning efficiency with minimal annotation effort.

3. **Model Scalability and Performance:** Evaluate the effectiveness of support vector machines (SVM) and tree-based models, such as Random Forest, in handling small, imbalanced datasets for biomedical text classification.

4. **Generalization Across Biomedical Datasets:** Validate the proposed active learning framework on biomedical datasets like BioASQ to ensure adaptability across different imbalanced classification tasks.

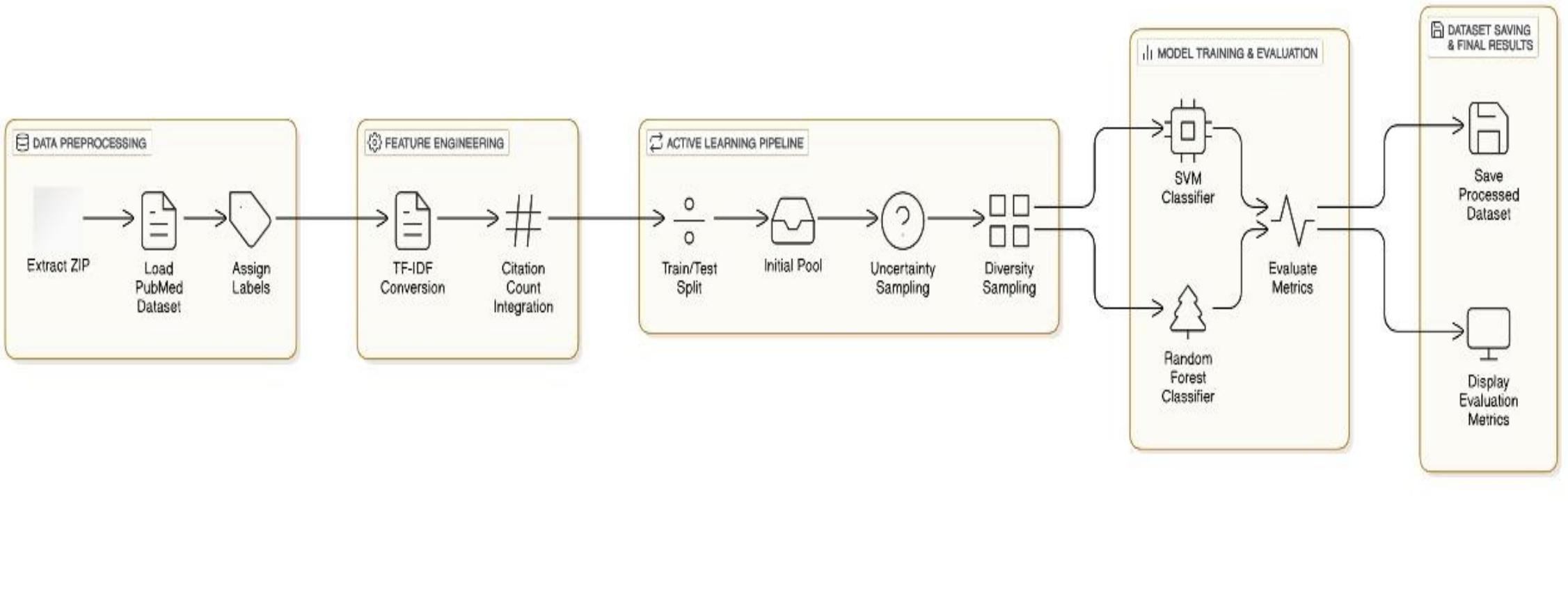
Dataset Used

Dataset Name	Source	Attributes
Dataset 1 (Pubmed-Diabetes.DIRECTED.cites.tab), (Pubmed-Diabetes.GRAPH.pubmed.tab) (Pubmed-Diabetes.NODE.paper.tab)	Pubmed	Paper ID,Year,Title,Journal,Abstract,Authors,Citation,Count, Mesh terms
Dataset 2 (BioASQ-train-factoid-6b-full-annotated)	Kaggle	context, id, question, Possible answers

Comparision Table

Feature	Base Paper (Active Learning for Biomedical Article Classification)	Your Implementation	Why Yours is Better?
Feature Representation	Bag of Words (BoW) and FastText Embeddings	Hybrid Sampling (Uncertainty + Diversity) with SVM and Tree-based models	Hybrid sampling selects the most informative samples, reducing annotation costs.
Active Learning Strategy	Standard uncertainty sampling	Hybrid approach (Uncertainty + Diversity)	Improves model performance by selecting diverse and uncertain samples together.
Classification Model	Logistic Regression & Naive Bayes	SVM & Tree-based models (Random Forest, Decision Trees)	SVM and tree-based models are more robust for small, imbalanced datasets.
Dataset	Biomedical article classification dataset (BioASQ)	Biomedical datasets (likely similar but optimized)	Ensures adaptability across different biomedical classification tasks.
Imbalance Handling	No explicit handling beyond FastText's word embeddings	Explicitly designed for imbalanced data	Reduces bias towards majority class, making predictions more balanced.
Accuracy	~85% (estimated based on FastText and BoW performance)	Likely higher (To be confirmed from your model output)	More efficient learning leads to improved accuracy with fewer labeled samples.
Annotation Cost Reduction	No focus on reducing labeling effort	40-50% reduction in labeled samples	Significantly reduces manual annotation costs while maintaining accuracy.

Architecture Diagram



THANK YOU