Harnessing the Power of Beta Scoring in Deep Active Learning for Multi-Label Text Classification

Wei Tan[1], Ngoc Dang Nguyen[1], Lan Du[1*], Wray Buntine[2]

[1]Faculty of IT, Monash University
[2]VinUniversity
wei.tan2@monash.edu, dan.nguyen2@monash.edu, lan.du@monash.edu, wray.buntine@vinuni.edu.vn

Abstract

Within the scope of natural language processing, the domain of multi-label text classification is uniquely challenging due to its expansive and uneven label distribution. The complexity deepens due to the demand for an extensive set of annotated data for training an advanced deep learning model, especially in specialized fields where the labeling task can be labor-intensive and often requires domain-specific knowledge. Addressing these challenges, our study introduces a novel deep active learning strategy, capitalizing on the Beta family of proper scoring rules within the Expected Loss Reduction framework. It computes the expected increase in scores using the Beta Scoring Rules, which are then transformed into sample vector representations. These vector representations guide the diverse selection of informative samples, directly linking this process to the model's expected proper score. Comprehensive evaluations across both synthetic and real datasets reveal our method's capability to often outperform established acquisition techniques in multi-label text classification, presenting encouraging outcomes across various architectural and dataset scenarios.

Introduction

Text classification is an essential task in natural language processing (NLP) with various applications, including sentiment analysis and topic classification (**??**). While deep learning has significantly improved multi-class text classification, yet its sub-task, multi-label text classification (MLTC), presents unique challenges. In MLTC, samples can be assigned multiple labels, navigating a vast, sparse label space with unresolved complexities (**??**). For every multi-label instance, the determination of appropriate labels adds to the intricacies and costs of annotation, making it even more resource-intensive than single-label learning. The increased demand for data becomes especially evident in specialized sectors, like the medical field, where domain-specific expertise and privacy considerations are crucial (**?**). Recently, active learning (AL) has been explored as a promising solution to MLTC's data scarcity issue, focusing on the annotation of the most informative instances. However, a comprehensive AL framework for MLTC, which robustly tackles imbalanced label distributions, has yet to be fully realized (**??**). In spite of these challenges, MLTC remains a critical task in NLP with diverse practical applications.

AL can alleviated the complexities MLTC, primarily due to its capacity to identify high-information samples, which subsequently optimizes human annotation and enhances model performance (**??**). Binary Relevance (BR) plays a key role in MLTC for its ability to segment multi-label tasks into individual binary classifications, facilitating the inclusion of advanced classifiers like deep neural networks and support vector machines (SVMs) (**?**). Inspired by BR's framework, earlier work in multi-label active learning (MLAL) adopted standard SVMs to establish binary classifiers tailored to individual labels (**???**). Contrary to the notion that BR overlooks label correlation, when paired with deep neural networks, label correlations are implicitly handled, addressing a critical aspect of multi-label classification (**?**). However, this focus on label correlation should not overshadow another crucial aspect of MLTC: data imbalance. The skewed distribution of labels in many real-world datasets affects instance selection strategies, potentially biasing models towards frequent labels (**?**). Therefore, recent studies, emphasize the need to explore the inherent properties of MLTC, especially label imbalance, as a pathway to enhance generalization performance, making it imperative to address this issue precisely (**?**).

In this paper, we propose a hybrid approach, named **Be**ta **S**coring **R**ules for Deep **A**ctive Learning (BESRA[1]) that incorporates the Beta family of proper scoring rules within the Expected Loss Reduction (ELR) framework, applied to MLTC. Unlike expected loss used in other MLAL methods, BESRA pinpoints informative samples involving both uncertainty and diversity, drawing on expected score changes within a predictive model governed by the Beta scoring function. Through extensive experimentation on diverse architectures (*e.g.,* TextCNN, TextRNN, BERT) and datasets (*e.g.,* Eurlex, RCV1, Bibtex, Delicious, Yahoo Health and TMC2007), we reveal that BESRA, even in the presence of varying label imbalances, consistently delivers robust performance, outperforming other AL methods. Our BESRA

---

*Corresponding author.

[1]Our implementation of BESRA can be downloaded from https://github.com/davidtw999/BESRA.

is adept handling of label imbalances sets a promising direction for future MLAL research and demonstrating robust performance in diverse data scenarios.

Related Works

AL is an iterative method that aims to reduce the cost and effort required for training high-quality prediction models by selecting the most informative instances from a pool of unlabeled data and querying an oracle for their labels (**??**). From the literature, AL studies typically focus on multi-class classification (MCC) problems with two common strategies, including uncertainty-based and diversity-based sampling (**?**). Uncertainty-based methods choose the instance that is most uncertain to label for the current trained classification model and quantify uncertainty using entropy or disagreement from ensembles (**?**). Diversity-based methods acquire informative and diverse samples using sample representations generated by pre-trained models (**?**). However, relying solely on either uncertainty or diversity may not be adequate to select informative samples for annotation, as demonstrated by previous studies (**??**). To overcome this challenge, hybrid methods balance both uncertainty and diversity to acquire the most informative samples. These methods often generate sample representations from the loss that affects the model's performance and includes predictive uncertainty (**??**). Hybrid methods have shown promise in enhancing AL for MCC by combining the benefits of uncertainty and diversity.

AL for multi-label classification (MLC) is an area of ongoing research, and current methods often involve transforming the multi-label problem into one or more single-label problems, allowing for the application of traditional single-label learning algorithms (**??**). In MLC, Binary Relevance is a widely used method that decomposes the problem into a set of binary classification problems, enabling instance selection decisions through the independent exploitation of binary classifiers (**?**). For example, **?** presents a strategy called maximum loss reduction with maximal confidence (MMC), which uses multi-class logistic regression to predict the number of labels for an unlabeled instance. Subsequently, it computes the expected loss associated with the most confident result of label prediction from support vector machine classifiers across all labels. To improve the MMC framework, **?** proposed a soft Hamming loss reduction (SHLR) by extending MMC for MLAL by introducing a major and an auxiliary learner. The two learners are designed to evolve together during AL, and queries are made based on their disagreement. The framework provides flexibility in the choice of learners and query criteria and also includes MMC, BinMin, and the random query algorithm as special cases (**??**). Recently, ADAPTIVE (**?**) and CVIRS (**?**) consider label dependencies in their uncertainty sampling. The CSRPE (**?**) method introduces cost-sensitive coding to manage varying error costs. Additionally, the GP-B2M (**?**) method deals with label sparsity and seeks to reveal label relationships using a Bayesian Bernoulli mixture of label clusters. However, all these studies do not explicitly tackle the issues of label imbalance, which are frequently observed in MLTC, as indicated in Table 2. Our AL method offers a fresh approach to addressing the imbalance challenges in MLTC,

displaying consistent performance across both synthetic and real datasets.

Beta Distribution-based Scoring Rules

ELR (**?**) quantifies the generalisation error that is expected to reduce when a new sample is added to the labeled pool L. Leveraging the insights from previous research, it uses an unlabeled pool U to represent potential test data, thereby quantifying the expected generalization error (**?**). Mean Objective Cost of Uncertainty (MOCU) (**?**) generalizes this approach and adds theoretical rigour. The Bayesian Estimate of Mean Proper Scores (BEMPS) improves these formulations by replacing error with proper scoring rules (**?**), leading to an AL strategy proven to converge (**?**). Scoring rules measure the quality of predictive distributions by rewarding calibrated predictive distributions. Different scoring rules, including Brier score, Logarithmic score, and Beta family (**?**), focus the quality measurement on different aspects of the model prediction. Unlike the Brier and Logarithmic scores, which are tailored for MCC, the Beta distribution inherently models variables limited between 0 and 1, making it more apt for BR in multi-label problems.

The optimization goal of our multi-label active learner is to label data that leads to the largest change in expected scores measured by the Beta family on the prediction of all binary classifiers. Let $P_T(x)$ be the (unknown) input distribution and $P_T(y|x)$ be the (unknown) conditional distribution on label sets $y \in \mathcal{Y}$, wherein $\mathcal{Y}$ is the set of all label sets. This distribution $P_T(y|x)$ is used by the so-called Bayes classifier, an ideal because it is unknown. We denote the predicted probability of $y|x$ given the current running model θ as $f_L(x, \theta) = P(y \mid \theta, x)$, and the predicted probability of $y|x$ under a Bayes optimal classifier as $f_L(x) = P(y \mid L, x) = \int_\theta P(y \mid \theta, x)P(\theta|L)d\theta$.

A "quality" score for the active learner is formulated as the expected difference Q_L between the score of the true probability estimates and a Bayes classifier after training on L (**?**), following for instance **?**, Sct. 4.1. However, the true conditional distribution $P_T(y \mid x)$ is unknown, so we make do with the current Bayesian estimate $f_L(x)$ and average the score differences across different models θ:

$$
\begin{aligned}
Q_L &= \mathbb{E}_{P_T(x)}\mathbb{E}_{P(\theta|L)}\Big[\mathbb{E}_{P(y|\theta,x)} \\
&\quad \big[S(f_L(x,\theta),y) - S(f_L(x),y)\big]\Big], \quad (1) \\
\Delta Q(x|L) &= Q_L - \mathbb{E}_{P(y|L,x)}\big[Q_{L+\{x,y\}}\big] \\
&= \mathbb{E}_{P(y|L,x)}\Big[\mathbb{E}_{P(x')P(y'|L,(x,y),x')} \\
&\quad \big[S(P(\cdot \mid L,(x,y),x'),y') - S(P(\cdot \mid L,x'),y')\big]\Big] \quad (2)
\end{aligned}
$$

Here S is a scoring function which scores a predictive probability distribution given the true label y. The active learner should evaluate Eq (1)- with L given by $L + \{x, y\}$ for each possible unlabeled data $x \in U$ to find the optimal query x^*. Since labels y of any new sample of data x are unknown, the $Q_{L+\{x,y\}}$ should be averaged over labels again using the Bayesian estimate. The formula to maximise for unlabeled sample x then becomes the difference, $\Delta Q(x|L)$, as in Eq (2), which can be further simplified (**?**, Equation 6).
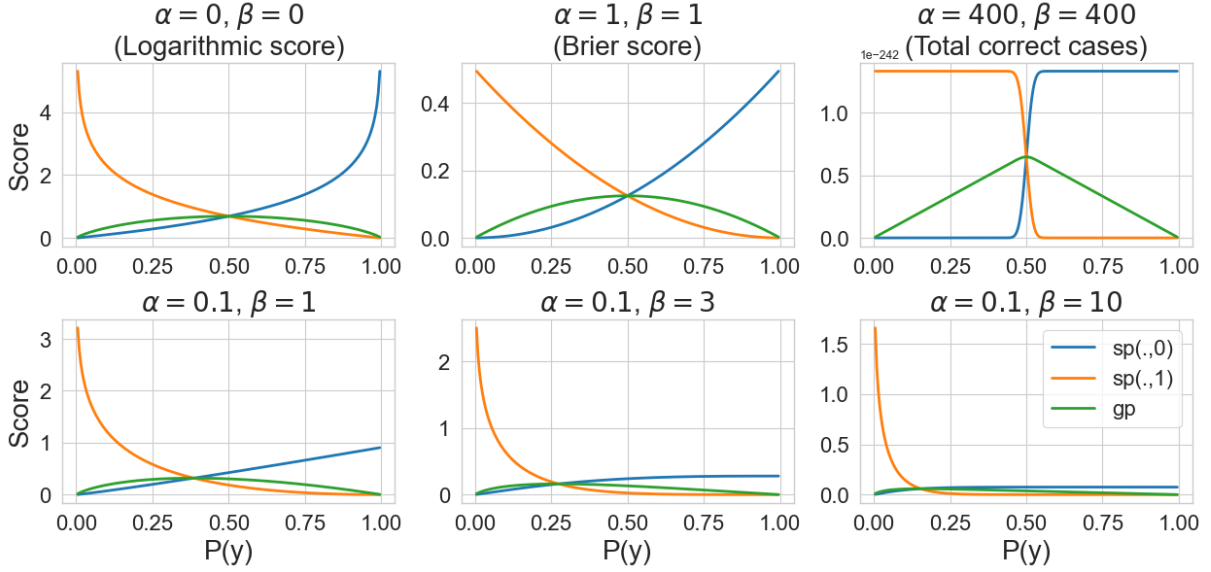
Figure 1: The graph depicts the Expected Score gp (green) and Scoring Functions from Eq (5) for the Beta family in blue (sp(., 0) when $y = 0$ and orange (sp(., 1) when $y = 1$). It covers six scenarios: three specific to Brier Score, Logarithmic Score, and total error approximations, and three emphasizing asymmetry with varied Beta values.

In the multi-label setting where there are K classes, and each class can be made a binary classification problem with $y_k \in \{0, 1\}$, the scoring function using BR is $S_{BR}(f_L(x), y) = \sum_{k=1}^{K} S^k(f_L(x), y_k)$ where $S^k(\cdot, \cdot)$ is the score for each binary class (**?**). Because of linearity, this works neatly with Eq (2).

For implementation, we follow the algorithm of BEMPS in (**?**) using our S_{BR} score. For this, $\mathbb{E}_{P_T(x)}$ is approximated as the average over a smaller estimation pool (sampled from U). Moreover, $\mathbb{E}_{P(\theta|L)}$ is approximated with an ensemble $\frac{1}{|\Theta|} \sum_{\theta \in \Theta}$ as is $\mathbb{E}_{P(y|L,x)}$ as $\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{E}_{P(y|\theta,x)}$. The Bayesian update required for $P(y' \mid L + \{x, y\}, x')$ is done by reweighting the ensemble:

$$P(y' \mid L, (x, y), x') = \sum_{\theta \in \Theta^E} P(y' \mid \theta, x') P(\theta \mid L, (x, y)) \quad (3)$$

$$P(\theta \mid L, (x, y)) \approx \frac{P(\theta \mid L) P(y|\theta, x)}{\sum_{\theta \in \Theta^E} P(\theta \mid L) P(y \mid \theta, x)} \quad (4)$$

Now the sum of strictly proper scores leads to a strictly proper score (**?**). So if $S^k(\cdot, \cdot)$ is a strictly proper score, then it follows that our score $S_{BR}(\cdot, \cdot)$ will be strictly proper, and therefore all the theoretical benefits ensue for our MLAL framework, such as provable convergence. We propose using the Beta family introduced by (**?**), a two-parameter family of proper scoring rules $\alpha, \beta > -1$ particularly useful for imbalanced classes or unequal costs.

$$S^k_{\alpha,\beta}(p, y_k) = \begin{cases} - \int_0^p c^\alpha (1 - c)^{\beta-1} dc \\ = -\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} I_p(\alpha + 1, \beta) \quad , y_k = 0 \\ - \int_p^1 c^{\alpha-1}(1 - c)^\beta dc \\ = -\frac{\Gamma(\alpha)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+1)} I_{1-p}(\beta + 1, \alpha) \quad , y_k = 1 \end{cases} \quad (5)$$

Here, $I_x(a, b)$ denotes the Incomplete Beta Function for $a, b > 0$, and the probability vector $P(y) = (p, 1 - p)$ given as p. If one of $a, b \leq 0$, an alternative evaluation is required.

The Beta family of scoring rules generalise the Logarithmic score and the Brier score used in BEMPS as follows (**?**): When $\alpha = \beta = 0$, Eq (5) reproduces the Logarithmic score and when $\alpha = \beta = 1$, it yields the Brier score. From this perspective, BESRA can be regarded as the generalization of BEMPS. When $\alpha = \beta \to \infty$, the scoring function becomes a step function, which approximates misclassification score. However, unlike the Brier and Logarithmic scores, which equally penalize both the low probability region (i.e., False Negatives (FN)) and the high probability region (i.e., False Positives (FP)) (as seen in Figure 1), the Beta family allows for a more tailored approach. This adaptability is especially beneficial in scenarios with specific demands, such as imbalanced multi-label classifications (see Table 2). Adjusting the α and β values, the Beta Scoring Rules can be adjusted to differentially penalize false positives and negatives. For example, by fixing $\alpha = 0.1$ and tweaking β, we can incline the penalties more towards FN predictions. Conversely, selecting the right β and α is critical and varies based on the problem, as highlighted in (**??**). Our research explores how varying these parameter values for BESRA impacts MLAL's performance in the Beta Parameters section below.

After incorporating the obtained samples into the labeled set, we proceed to retrain the multi-label classifier (e.g., TextCNN, TextRNN and BERT) using randomly initialized parameters. This retraining occurs after each iteration of acquisition and is carried out using the binary cross-entropy loss (BCE), a commonly employed loss function in multi-label learning (**?**). Specifically, given a specific data point x, we compute the probability of x being affiliated with class k, $\forall k \in \{1, 2, \ldots, K\}$ as $p(y_k = 1|x) = \frac{1}{1+\exp(f_k(x))}$, where f_k is the Binary Relevance associated with class k. BESRA's Algorithm 1 uses Bayes theorem to estimate $P(\theta|L, (x, y_k))$

Algorithm 1: Beta Scoring Rules For Deep Active Learning

Require: initial unlabeled pool U, initial labeled pool L, model ensemble $\Theta^E = \{\theta_1, ..., \theta_E\}$ built from L by retraining, pre-computed values for $P(y_k \mid \theta, x)$ and $P(y_k \mid L, x)$ for all relevant y and x, acquire batch size B, acquire batch set $\mathcal{A}$, estimation pool X, number of acquisition iterations N, number of classes K.

1: Initialize: $i = 0, L_0 \leftarrow L, U_0 \leftarrow U$
2: **while** $i < N$ **do**
3:    $vec_{x,x'} = 0$
4:    **for** $x \in U_i$, $x' \in X$ **do**
5:      **for** $k \in K, y_k \in \{0, 1\}$ **do**
6:       Compute $P(\theta|L, (x, y_k))$ with Eq (4) for for each y_k
7:       Compute $p(y'_k|L + \{x, y'_k\}, x')$ with Eq (3) for each y'_k
8:      **end for**
9:      $vec_{x,x'} += \Delta Q(x|L, x')$, where $\Delta Q(x|L, x')$ is computed by Eq (2) with the beta scoring function defined in Eq (5).
10:    **end for**
11:    $\mathcal{A} = \emptyset$
12:    $centroids$ = k-Means centers $(vec_{x \in U_i}, B)$
13:    **for** $c \in centroids$ **do**
14:      $\mathcal{A} \cup= \{ \arg\min_{x \in U_i} ||c - vec_x|| \}$
15:    **end for**
16:    $L_{i+1} \leftarrow L_i \cup \mathcal{A}$
17:    $U_{i+1} \leftarrow U_i \setminus \mathcal{A}$
18: **end while**

Dataset	#Document Train/Test	#Vocabulary /#Label	MeanIR Train/Test
RCV1-10	1,200/600	23,759/10	10/50
RCV1-50	1,200/600	23,117/10	50/50
RCV1-200	1,200/600	21,074/10	200/50
RCV1-400	1,200/600	34,464/10	400/50

Table 1: Five datasets (including the test set) are created using the MeanIR values that represent the imbalance level of the multi-label datasets. Higher values suggest increased sparsity and imbalance.

from $P(\theta|L)$, leveraging a predefined estimation pool X, a randomly selected subset from unlabeled pool, for estimating expected values $\mathbb{E}_{P(x')}[\cdot]$ in Eq (2). Each query is now represented as a vector of expected scored changes, i.e., $vec_{x,x'}$. These vectors are then organized using k-Means clustering, and from each cluster, only the data points closest to the centroid are selected for precision and representation.

Experiments

We evaluated BESRA's impact on synthetic and real-world datasets, particularly studying the influence of Beta Scoring Rules on different imbalance levels in synthetic data. Further, we benchmarked BESRA against state-of-the-art (SOTA) AL methods, affirming BESRA's efficacy in MLTC.

Dataset	#Document Train/Test	#Vocabulary /#Label	MeanIR Train/Test
Bibtex	5,916/1,479	1,836/159	12/15
TMC2007	22,876/5,720	497/25	22/23
Delicious	12,872/3,219	501/982	72/73
RCV1	20,833/7,965	119,475/102	285/152
Eurlex	15,470/3,868	5,000/201	426/164
Yahoo	7,364/1,841	30,587/32	656/162

Table 2: Six benchmark datasets with their corresponding imbalance level MeanIR statistics.

Datasets

Synthetic Datasets. We created five synthetic datasets, each consisting of ten labels and associated with a unique mean imbalance ratio (MeanIR), a measure of the average level of imbalance in the multi-label dataset (**?**), as outlined in Table 1. Four of these datasets, aligned with MeanIR values of $(10, 50, 200, 400)$, served training purposes. Additionally, a consistent test set with a MeanIR of 50 was established for standardized evaluations (**?**). To make the generation process for the synthetic dataset controllable, we reduced the label space to ten (i.e., $K = 10$) by selecting the large dataset RCV1, and sampled instances based on the MeanIR metric provided by **?**. In contrast, label cardinality measures the average labels per instance and label density evaluates the instance-label proportion. However, neither offers a holistic perspective on the overall label imbalance. As our main objective is to develop a general AL framework for MLTC that can handle the label imbalance issue, These synthetic datasets are crucial for evaluating the performance of AL methods with the Beta Proper Scores under varying levels of label imbalance. The experiments used a greedy search to identify the optimal parameters for the Beta scoring function ($\alpha = 0.1, \beta = 3$) compared to the Brier score ($\alpha = 1, \beta = 1$) and the Logarithmic score ($\alpha = 0, \beta = 0$). By systematically manipulating the Alpha and Beta values, the study evaluates BESRA's robustness and quantifies its influence across different magnitudes of dataset imbalance.

Real Datasets. We have selected six diverse benchmark multi-label text datasets (MLTDs): Eulex (**?**); RCV1 (**?**), comprising Reuters newswire stories; Bibtex (**?**), which contains bibliographic records in BibTeX format; Delicious (**?**), capturing user bookmarking data; Yahoo (**?**), detailing health-related user discussions; and TMC2007 (**?**), a collection of texts from varied domains. These datasets' label spaces vary significantly, ranging from 25 to 982 labels. We used the MeanIR statistics to measure the imbalance level in each. For example, Bibtex possesses the lowest MeanIR of the training set, meaning it's less imbalanced than the others. In contrast, Yahoo has the highest MeanIR, indicating it's the most imbalanced dataset among those we considered. To ensure data integrity, we have kept both train and test sets in their original sizes. Table 2 offers a summarized overview of these datasets' key characteristics for in-depth analysis and research.
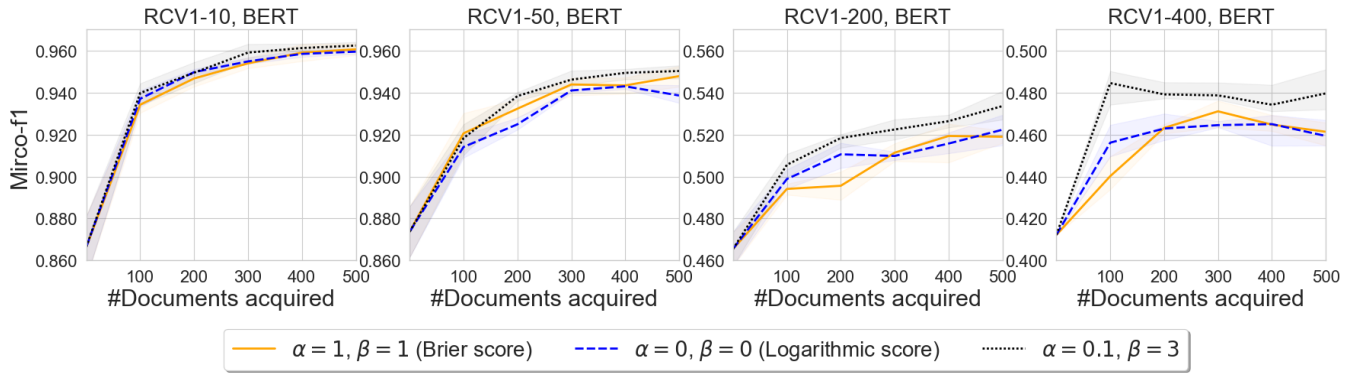
Figure 2: The average Micro F1-score of AL models with acquisition size 100 on BERT, which were run with 5 different random seeds on various synthetic datasets.

Model Architectures

In our experiments, we use TextCNN (**?**), TextRNN (**?**), and BERT (**?**) as the backbone classifier. After each AL iteration, we fine-tune these models using random re-initialization (**?**) for enhanced efficacya method found to be more effective than incremental fine-tuning with new samples (**?**). The fine-tuning process, executed on RTX3090 GPUs, sets a maximum sequence length of 256, runs for up to 80 epochs, and uses the AdamW optimizer with a 1e-5 learning rate (**???**). Within the AL setup, the initial training and validation sets are composed of 100 and 1000 samples from the training set. To ensure robustness, each AL method iterates five times on each dataset with the random seeds, sampling 100 instances in each acquisition. By incorporating a deep ensembles method (**?**), we compute predictive distributions $P(y|L, x)$ using five ensemble models. Performance was assessed using metrics like micro-F1, macro-F1, precision, recall, precision@5, and recall@5, though the primary paper emphasizes the micro F1-score, and further metrics' results can be found in the Appendix.

Active Learning Baselines

We compared BESRA with various AL methods, including a random baseline. For each method, we adapted its strategies to the MLTDs and using a consistent backbone classifier for all. Among the methods examined, the **MMC** method by **?** selects instances by maximizing loss reduction; **ADAPTIVE** by **?** chooses instances by balancing prediction uncertainty with label cardinality inconsistency; **AUDI** emphasizes label ranking and adjusted cardinality inconsistency, targeting uncertainty and diversity (**?**); **CVIRS** (**?**) leverages label vector inconsistency and score rankings to decide on instance labeling; **SHLR** by **?** employs an auxiliary learner to determine instance choices based on disagreement levels; **CSRPE**, as introduced by **?**, extends one-versus-one coding to a cost-sensitive approach, using code-bits to inform instance weights and improve performance; finally, **GPB2M** (**?**) combines a Gaussian Process with a Bayesian Bernoulli mixture to capture label correlations for instance acquisition.

Results

Synthetic Datasets. Figure 2 illustrates the micro F1-scores for three methods. Both the Brier ($\alpha = 1$, $\beta = 1$) and Logarithmic score ($\alpha = 0$, $\beta = 0$) equally penalize FNs and FPs. In contrast, the Beta score method with parameters ($\alpha = 0.1$, $\beta = 3$) emphasizes a more substantial penalty for FN predictions, leading to an enhanced AL performance. Average results from five AL experiments, with a 95% bootstrapped error band, underscore the Beta score's superior performance, especially on high imbalance datasets like RCV1-200 and RCV1-400. Although Brier shows a slight edge over the Logarithmic score in the RCV1-50 scenario, their differences diminish in other scenarios. Specifically, by optimizing for FN predictions, the Beta Proper Scores ($\alpha = 0.1$ and $\beta = 3$) have showcased their efficacy in handling varied imbalance distributions, especially as the imbalance ratio increases. However, these parameter values can be adjusted for specific applications based on desired scoring criteria, as discussed by (**??**). Further insights on this topic, including an ablation study and detailed discussion, are available in subsequent sections, with a more exhaustive analysis presented in the Appendix.

Real dataset. Figure 3 shows a performance comparison between BESRA and the baseline methods across six MLTDs, as listed in Table 2. We use the Beta score with $\alpha = 0.1$ and $\beta = 3$ due to its superior performance compared to the Brier score and Logarithmic score. Remarkably, BESRA consistently outperforms other AL methods in all acquisition scenarios, regardless of the text domains. Among the evaluated methods, ADAPTIVE has the highest overall performance based on the micro F1-score. It consistently demonstrated strong performance across all datasets, showing its effectiveness in MLTC AL. Asides from ADAPTIVE, AUDI also provided competitive performance on all datasets except for Yahoo (health), suggesting its potential in specific domains. Conversely, CVIRS and GPB2M exhibited inconsistent performance across the datasets, struggling particularly in EURLEX, RCV1, and TMC2007. CSRPE method yielded inadequate results in the Bibtex and Delicious datasets, indicating a lack of effectiveness in those specific contexts. MMC emerged as the least effective method
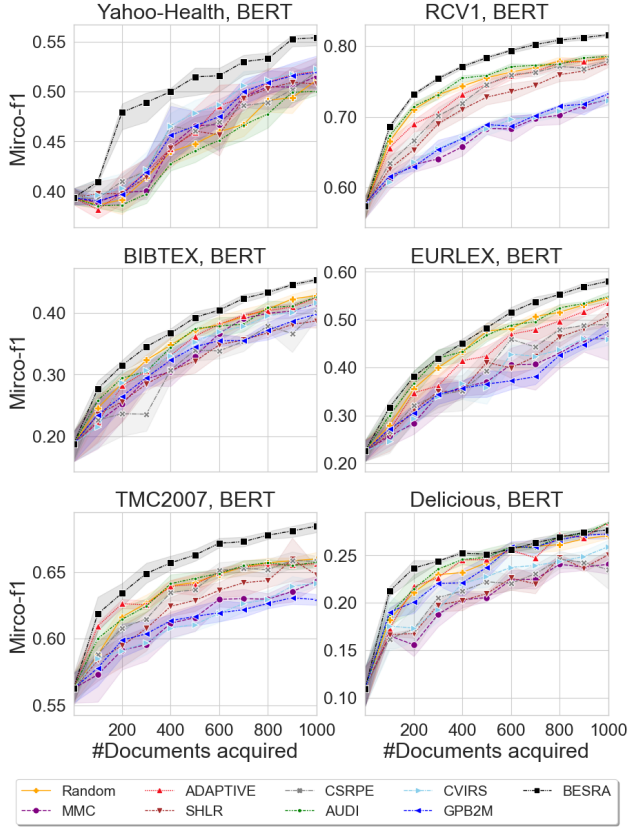
Figure 3: The average Micro F1-score of AL models with acquisition size 100 on BERT, which were run with 5 different random seeds on various datasets.
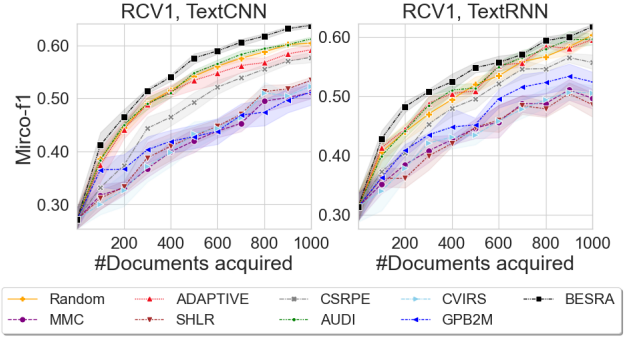

Figure 4: The average Micro F1-score of AL models with acquisition size 100 on TextCNN and TextRNN, which were run with 5 different random seeds on RCV1.
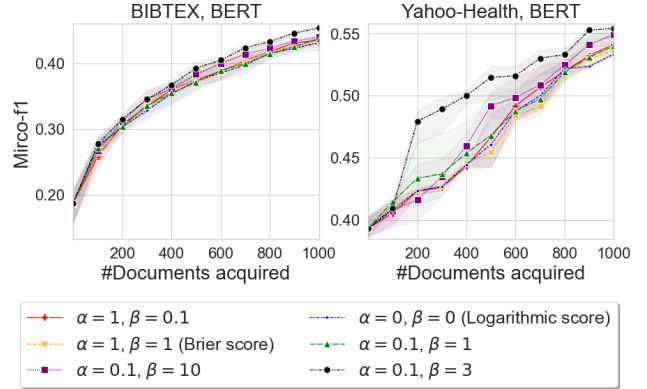

Figure 5: The average Micro F1-score of AL models with acquisition size 100 on BERT, which were run with 5 different random seeds on for Bibtex and Yahoo (health).

across all datasets except for Yahoo (health). In addition, MMC and SHRL showed similar performance, as both methods focused on measuring the expected loss of the model. It is worth highlighting that SHRL showcased better results than MMC specifically in the RCV1 and TMC2007 datasets, this could be due to SHLR being the improvement of MMC by introducing the soft Hamming loss (**?**) as an alternative to the mean of the expected loss used in MMC. GPB2M, the most recent work in ML AL, demonstrated superior performance in the Yahoo (health) and Delicious datasets, highlighting its efficacy in those specific contexts. However, the robustness of this AL strategy raises a significant concern, as its performance exhibits variability not only across diverse datasets but also among various model architectures. For a more comprehensive analysis, including additional metrics such as macro-F1, precision and recall, please refer to the Appendix.

Ablation Study

To assess the effectiveness and generalizability of BESRA, we conducted a comprehensive evaluation of three well-established neural network architectures commonly used in MLTC tasks: TextCNN, TextRNN, and BERT. These architectures have played a significant role in multilabel learn-

ing. Additionally, we investigated the impact of different Alpha and Beta values on the model's performance. Our study aimed to answer two key research questions: **(i)** Does the performance advantage offered by BESRA generalise across diverse models and architectures? **(ii)** How do the Alpha and Beta values influence the scoring mechanism and subsequently affect the model's performance across real-world datasets with the varied imbalance level? By addressing these questions, we aimed to gain insights of BESRA and its potential application in various MLTC tasks.

Model Generalizability

The evaluation of BESRA is being conducted in two other models to test its generalizability. As depicted in Figure 4, BESRA consistently outperforms other baselines across TextCNN and TextRNN tested on the RCV1 dataset. Additionally, BESRA demonstrates exceptional performance compared to these models on five additional datasets included in the Appendix. These results validate that BESRA demonstrates effectiveness across a diverse range of pre-trained language models, highlighting its applicability irrespective of the specific model architecture applied.
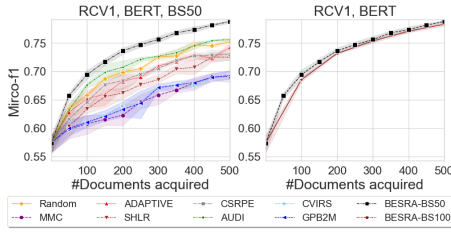
Figure 6: Left: Learning curves of ALs with batch size 50 on RCV1. Right: Learning curves for BESRA with batch sizes $B \in \{50, 100\}$ on RCV1. All results were run with 5 different random seeds

Batch Size

Figure 6 indicate that BESRA consistently outperforms other ALs when using a batch size of 50. Additionally, during the early stages of acquisition, BESRA performs more effectively with smaller batch sizes (50) than with larger batch sizes (100), aligning with the results reported in (**??**). We hypothesise that at the early acquisition stages, due to the lack of knowledge, the model trained with a limited number of samples has high uncertainty and lacks calibration, thus acquiring a large batch of samples can result in noise. Additional experiments conducted on other MLTDs can be found in the Appendix.

Beta Parameters

In this subsection, we delve into the implications of varying Beta scores on the performance of AL, with a particular emphasis on how these scores influence the penalization behaviour of the active learner. We evaluated a range of α and β values, associated with distinct Beta scores as shown in Eq (5). Our evaluations spanned three representative datasets, ranging from relatively balanced (i.e., BIB-TEX) to highly imbalanced (i.e., Yahoo). To gain a comprehensive understanding of how different α and β values impact the outcomes, we considered several scoring methods including the Brier score and Logarithmic score (which provide equal penalization), alongside four distinct scenarios of Beta score. These scenarios include (1) mild penalization on False Positives (FPs) where $\alpha = 1$ and $\beta = 0.1$, (2) light penalization on False Negatives (FNs) with $\alpha = 0.1$ and $\beta = 1$, (3) moderate penalization on FNs defined by $\alpha = 0.1$ and $\beta = 3$, and finally (4) stringent penalization on FNs denoted by $\alpha = 0.1$ and $\beta = 10$. Insights from Figure 5 reveal that the Brier and Logarithmic scores, as well as the scenario with $\alpha = 1$ and $\beta = 0.1$, consistently underperform across various imbalance levels within MLTDs. Such results are consistent with our prior expectations, considering the inherent label imbalance challenges. A significant reason for this underperformance is the equal penalization rendered to both FN and FP outcomes by the Brier and Logarithmic scores. Additionally, the specific scenario of $\alpha = 1$ and $\beta = 0.1$ tends to disproportionately penalize FPs, thereby dampening performance. When evaluating the effect of α and β values focused on penalizing FNs, we note that while light penalization settings (i.e., $\alpha = 0.1$,

$\beta = 1$) have a negligible impact on enhancing the active learner's effectiveness, the more stringent configuration of $\alpha = 0.1$ and $\beta = 10$ offers notable improvements, especially in highly imbalanced datasets. However it does not necessarily culminate in the optimal active learner. Instead, a moderate penalization strategy with $\alpha = 0.1$ and $\beta = 3$ consistently stands out as the most effective across MLTDs.

Conclusion

We have introduced BESRA, a novel acquisition strategy for MLAL. This generalizes the recently published BEMPS using the Beta family of proper scoring rules, which allow customizable asymmetric scoring rules that effectively address the challenges such as imbalanced data associated with multi-label learning. Moreover, by our methodical construction, the use of BESRA provably converges to optimal solutions. Through empirical studies conducted on synthetic and real-world datasets, we have demonstrated the effectiveness of BESRA in acquiring highly informative samples for multi-label active learning, consistently surpassing seven existing acquisition strategies. This finding highlights the crucial role of Beta Scoring Rules and their great potential for AL with tailored acquisition strategies. Future research can further explore combinations of Alpha and Beta values for specific datasets, addressing a current limitation of BESRA. Quia quas officia labore ex ad magni id voluptatem ut, quia sit odio ab at aperiam nesciunt in omnis quam ipsa porro, asperiores laudantium dignissimos beatae a commodi, ex quasi repellat modi dignissimos veritatis explicabo necessitatibus ut ipsa, quaerat quam pariatur velit voluptatum explicabo facere?Nulla minus excepturi quo provident totam consequuntur cumque nostrum facilis, laudantium sapiente praesentium architecto ipsum eius veritatis recusandae in, et nobis sed repellat eaque ducimus, quisquam quo corrupti harum officia fugit?Fuga numquam consequatur, tempore aliquid ducimus?Magni cupiditate doloremque eius adipisci aut assumenda quo corrupti modi, voluptatem commodi fugiat architecto minus voluptas, dolores iusto corrupti distinctio perspiciatis quaerat quis illo sunt vel tenetur, ad dolorem incidunt sed dolorum quos saepe cum suscipit ab quisquam ipsa.Quos dicta similique illo tempora omnis facere eveniet explicabo ipsum, similique rerum in obcaecati suscipit.Qui eius pariatur dolorem facere fugiat, corporis eos ad eum, eligendi dignissimos dolorum totam.Quisquam ipsa reiciendis nostrum modi blanditiis officia unde nobis explicabo, eius nisi ab debitis odit magni omnis tempore, aliquid aut modi iste non ducimus, aperiam sint tempora recusandae quidem dolorum distinctio minima laboriosam incidunt?Facilis sit ipsam iure fugiat perferendis cum, error illo quis est assumenda, ad quis cupiditate ut quos itaque dolorum asperiores cumque maxime incidunt.Iste beatae debitis aut earum dignissimos sed, exercitationem fugit sunt iusto iste velit sequi a voluptatum, hic sapiente repudiandae sunt quas doloremque aliquid magni corrupti cupiditate ipsa, explicabo quaerat laboriosam enim tenetur et?Quasi temporibus dolorem nobis nostrum consequatur blanditiis vero provident aperiam, minus cum officia architecto est corporis ut qui molestiae recusandae, sunt possimus illum tempora sed sequi voluptate vitae obcaecati doloremque