Addressing Class Imbalance in Active Learning via Approximate Nearest Neighbour Methods

Lian Fichman {lianfi@campus.technion.ac.il}, Idan Horowitz {idanhorowitz@campus.technion.ac.il}, Yarden Kamienney {yardenk@campus.technion.ac.il}, Carmel Soceanu {carmel.so@campus.technion.ac.il}, Technion

September 2024

Abstract

Active Learning (AL) has been recognized for its ability to optimize model training by selectively querying the most informative data points. However, its efficacy on imbalanced datasets, particularly in high-dimensional, multiclass classification settings, remains underexplored. This research proposes an intergration of Approximate Nearest Neighbor (ANN) sampling methods into an AL pipeline, as well non-random initialisation for the AL initial train set. The study employs various ANN indices, including FAISS and ANNOY, to efficiently handle high-dimensional language embeddings. We evaluate the impact of our different suggested AL sampling strategies - distance-based and cluster-based sampling, on multiple imbalanced textual datasets. Results indicate that while both distance and cluster-based methods show potential and improve on random sampling, they are surpassed by traditional margin sampling. Nonetheless, the integration of ANN methods with AL demonstrates promise, particularly in enhancing sampling efficiency and improving class coverage.

1 Introduction

Active Learning (AL) allows a model to query the most informative data points from a pool of unlabelled data. This enables the model to optimize the learning process and improve classification performance. While AL methods have demonstrated success, (Ertekin et al., 2007), particularly in optimising learning times in large datasets (Rittler & Chaudhuri, 2023), the field still has avenues that have not been fully explored.

Imbalanced data describes a scenario in which one or several classes are over-represented at the expense of other, underrepresented classes. This leads to learning failures, as the models to not have enough access to the under-represented class in order to make effective predictions, or place too much emphasis on the majority class. This problem is notoriously difficult to solve, as traditional approaches such as SMOTE (Chawla et al., 2002) have been shown to lack effectiveness, especially in high-dimensional datasets (Ali et al., 2013), where the underlying distribution is misrepresented owing to the resampling.

Intuitively, Active Learning should address some of the challenges encountered in imbalanced data - if the model could focus on varying samples across the classes more evenly, this might improve learning accuracy. Indeed, success has been seen in integrating AL on imbalanced datasets (Ertekin et al., 2007), (Aggarwal et al., 2021), but the success does not include comprehensive guarantees, particularly within a multiclass classification setting. Most standard AL sampling techniques are difficult to scale onto high-dimension datasets, and the runtimes become increasing less feasible.

Using a nearest-neighbours method in sampling was explored by Hu et al. (2009), and further advanced with SEALS (Coleman et al., 2021). Additionally, improving the coverage of the initial training set as opposed to a random initialization by means of a clustering approach, was shown to improve learning efficiency by Joshi et al. (2012), and a more complex version of this, including the aggregation of various distance and similarity metrics within each cluster to aid sampling obtained impressive results for Genossar et al. (2023). Nonetheless, much of the current literature focuses on low-dimensional image data, and does not harness the power and efficiency of indexing methods with high-dimensional language vectors. With the exception of Genossar et al. (2023), there is little focus on an aggregation of methods, or a combination of approaches. Thus, this paper proposes an ANN-based approach to improving classification in an Active Learning pipeline for an imbalanced multiclass setting on large language vectors.

2 Method

2.1 Data

2.1.1 Dataset Selection

In order to find a suitable imbalanced dataset, an Active Learning pipeline using random sampling was run on 3 different datasets, assessing the effectiveness of the AL method and to obtain a baseline performance

measure. The Skin Cancer ISIC image dataset was used for a binary cancer-prediction classification task, while the IMDb Reviews Dataset was used for a binary sentiment classification task on the embedded vectors of textual reviews. The former dataset had a 2:1 ratio in terms of class imbalance, while for the latter a similar ratio was obtained by dropping half of one class[1]. In both cases, the random selection based AL pipeline obtained a very high classification accuracy[2], and thus the IMDb Movies Dataset was chosen, for which the random selection AL method in the task of predicting the film genre from the textual description of the film, in a multiclass classification setting performed very poorly[3]. Additionally, the class imbalance in this dataset was extreme, owing to the 20 different genres and an strong representation of Drama films, as depicted in Figure 1. Thus, this dataset was selected to

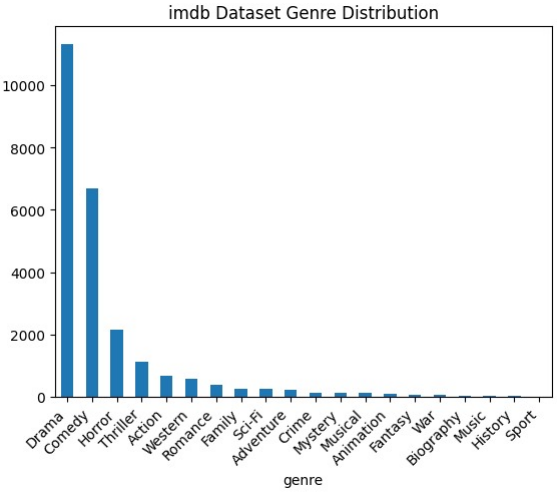


Figure 1: Label Distribution

serve as the basis of the research, exploration and assessment of different methods. Further datasets used for evaluation are described in 2.1.3. At this stage, the language vectors were embedded via the TF-IDF vectorizing method, owing to its simplicity and ease of use for an initial exploration. See 2.1.2 for the embeddings used in the research and evaluation states

2.1.2 Data Preprocessing & Embedding

The preprocessing all datasets, including those used for evaluation (see 2.1.3 below), was straightforward. Null values were removed, and since the classification task was to focus only on the textual vector embeddings, all data relating to each sample was also dropped, barring the target label, which would act as a label. The text fields were processed by removing punctuation and stopwords, the latter as defined by the Natural Language Toolkit's (NLTK) stopwords package. To embed the textual descriptions used as features, 3 pre-trained transformer-based embeddings were assessed, all obtained from HuggingFace's sentence transformers package. The embeddings are obtained from the mp-net, RoBERTa, and BERT transformers, which were chosen for their all-round, general-purpose embeddings that are well adapted to a variety of tasks - a fitting choice for a specific research project with potential future wider applications, not requiring further fine-tuning and able to serve as an effective benchmark. All embeddings resulting from the models are of dimension 768.[4] The impact of embeddings on performance was minor, and insignificant given that the same embedding is used for all benchmarks. The MP-Net model was ultimately chosen for the task.

For the IMDb research dataset, nearly 2/3 of samples were multi-label entries, with more than one genre. In order to reduce runtime and limit task difficulty that is not as a result of the label imbalance[5], these examples were dropped, allowing for a focus on only single-label entries. The genres were encoded using a standard numerical encoding[6] and the resulting dataset contained approximately 25000 labelled movies.

2.1.3 Evaluation Datasets

After the research phase performed on the IMDb movies dataset described in 2.1.1, the methods detailed in 2.2 and 2.3 were evaluated on 4 additional datasets. This is to ensure that the results obtained are not specific to a single domain, but hold up through a wide range of tasks, and to eliminate any issues involved in the hyper-focus on a specific dataset. All dataset followed the same paradigm of in vector embeddings representing a body of text comprises the sole set of features in a multiclass classification task. The datasets were processed and embedding in an identical manner to the IMDb dataset, described in 2.1.2, barring the dropping of the multi-label entries, as all datasets except one were singly labelled. The datasets were all obtained from Kaggle and are described as follows. The arXiv Paper Abstracts dataset involves classifying academic papers into subject areas based on the paper's abstract. The Legal Citations dataset requires predicting the outcome of the citation of a legal case based on the case description. The E-Commerce Text dataset entails the prediction of a product's category from its title on an E-commerce

[1] The original dataset was perfectly balanced

[2] 0.92 and 0.84 respectively, after 10 iterations. Note that although recall is a more suitable method for assessing imbalanced datsets, a 2:1 imbalance ratio means that a 'dumb' classfier only picking the majority class would still obtain an accuracy of only 0.67. Therefore, these results still indicate an satisfactory classifier performance.

[3] In multiple attempts, neither classification accuracy nor a weighted F1 score could cross 0.65

[4] The Paraphrase Mini LM model was also assessed, but its output dimension is smaller, thus its embeddings less informative or powerful.

[5] Initial results in a multi-label setting yielded very poor results, with an optimal micro-weighted F1 score of 0.32 after 20 iterations.

[6] Although this is not necessarily the most effective encoding for optimal learning accuracy, this is not the primary focus of the research, and maintaining it as the standard across all benchmark pipelines is sufficient.

platform, while the Kashnitsky dataset contained the same task, but based on the product review.[7] The label distribution of these datasets can be found in the Appendix.

2.2 ANN Methods

The selection of an index for the language vectors was performed by comparing various methods and distance metrics for both runtime and performance. The index chosen to move forward into the Active Learning pipeline would demonstrate a balance of both efficiency and accuracy on the selected dataset. Owing to the high-dimension of the language vectors, the ground truth value was defined by cosine similarity between the vectors, and ANN effectiveness by the recall score at k, where k is a varying parameter representing the number of nearest neighbours. Indexing methods were run through two packages, FAISS, and ANNOY. While the former is optimized for high-performance, GPU-accelerated ANN, the latter is a memory-efficient, tree-based method that builds multiple random projection trees to perform fast approximate searches, designed primarily for disk-based large datasets with a focus on minimizing memory usage. Although the ANNOY package does not support more advanced methods such as Inverted File (IVF) and Locally Sensitive Hashing (LSH), these were still included in the analysis. From FAISS, flat indices using L2 distance, inner product and cosine similarity were all assessed, with the latter also being used to compare the influence of a GPU and multithreaded execution on runtime. FAISS does not have a native cosine similarity measure - a normalized inner product index was used to obtain this metric. Non-flat indices were LSH, HNSW[8] and IVF using both an L2 distance and cosine similarity. From ANNOY, distance metrics assessed were the same as those used for flat FAISS indices, while cosine similarity was used to measure the influence of an increased number of trees[9] on recall and runtime. Full results are reported in 3.1.

2.3 Active Learning Methods

The primary variation between Active Learning pipelines is the selection criteria, which defines how the model samples the most informative data points at each iteration. Here, the approaches explored can be separated into three categories - Distance-based Sampling strategies, which focus on utilizing the distance[10] between vectors as a selection criterion. Cluster-based Sampling strategies, group together the vectors and use metrics relating to each cluster as a means of selection, similar to the Battleship approach detailed by Genossar et al. (2023). Each strategy and the varying sub-approaches that were tested are described below. The third method, Non-Random Initialization, does not focus on the sampling, but rather specifically selects an initial training pool instead of randomly assigning vectors. Integrations of all strategies were also tested. Results are detailed in 3.2.

2.3.1 Distance-based Sampling

All distance-based methods involved performing a search over all vectors within the available pool of unlabelled data and computing the distance from them to all vectors in the current train set. For example, with m vectors in the current train set, and n vectors in the available pool, a search for the m nearest neighbours is performed on each of the n pool vectors, resulting in a distance matrix of $n \times m$. At every iteration, the samples selected for the training set from the available pool were those furthest from the train set vectors, defined by either the mean distance from all vectors, the maximum distance to any training vector, or the minimum distance to any training vector. This aims to select samples that are the most different from those in the current training set, thereby obtaining instances from the minority class at an early stage of training and improving train-set diversity. Experiments using a Flat Cosine-Similarity based index and an LSH index with $nbits = 16$ were both evaluated.

2.3.2 Cluster-based Sampling

Cluster-based methods focus on clustering the data around centroids using a FAISS-based vector index method, and sampling for the train set from the available pool on a cluster-by-cluster basis. Selection from each cluster occurred via 3 methods, each which was obtained through a different metric and was given an equal representation in the Active Learning budget. The first metric sampled the nearest points to the cluster centroid, and the second chose the most diverse points within the cluster, where diversity is defined by points with the highest mean distance from all other points in the cluster. The third measure utilized a traditional uncertainty sampling method on each cluster, in which the uncertainty used is the Least Confident Sampling, defined as: $uncertainty = 1 - \max(probs)$.[11]
Clustering was run natively through both FAISS and SciKit-learn's kmeans package[12], for a maximum of 100 iterations before convergence, on each round of the Active Learning process. Additionally, a non-random cluster initialisation method[13] was explored

[7] This dataset was intended for hierarchical classification - in this case, only the primary (level 1) label was used.

[8] Hierarchical navigable small world.

[9] 100 instead of 10, which was the standard used for all metrics.

[10] Distance in this case can refer to any metric used to measure the relationship between two vectors, be it Euclidean (L2) Distance, Cosine Similarity or Inner Product. However, as justified in 2.2, most instances here refer to cosine similarity.

[11] Note, this apporach does not always guarantee that the full budget is utilized - in the event that a point from a cluster is selected more than once, it was not duplicated, and added singulalry to the train set.

[12] There were no appreciable differences between the two.

[13] Note that this differents from the Non-Random Initialisation of the Train Set in the Active Learning, described in 2.3.3

- for each movie genre, an LLM was queried[14] to obtain a sentence best describing the genre. This sentence was then embedded in the same manner as the data vectors, and were used as initial centroids for the clustering method. Furthermore, a dynamic clustering method was explored, in which each iteration expands the number of clusters. These methods were aimed at improving the effectiveness of the clustering, thereby enabling a more efficient sampling process. Evaluation results of the clustering itself can be found in the Appendix.

2.3.3 Non-Random Initialization

The goal of Non-random Initialization was to increase the coverage over the dataset from the outset, generating an initial train set that is more representative of all classes and reducing the impact of the label imbalance. This is in contrast to standard AL techniques, which randomly select the initial points. Two initialization techniques were employed - the first was simply ensuring at least one example from each class is represented in the train set[15]. The second is inspired by Joshi et al. (2012) and makes use of clustering, creating a number of clusters that matches the number of classes, and sampling at least one example from each cluster.

2.3.4 Pipeline & Model Integration

An active learning pipeline was built with 250 samples in the initial train set, a budget of 500 samples to select from the unlabelled pool at every iteration, and was run for 20 iterations. Since model development was not the purpose of this project, several simple models were experimented with for integration into the pipeline to perform the classification task, with the most crucial point being the consistent use of the same model for all baselines and comparisons. The model used were a Random Forest classifier, an SVM classifier, which historically has shown to be preferred and effective in an Active Learning setting, and a simple feed-forward neural network with 2 hidden linear layers and non-linear activation functions. Ultimately, the SVM classifier Stochastic Gradient Descent training defined to minimise the log-loss function was selected, as it outperformed the other two.

3 Experiments & Results

In the following section, the evaluation metrics refereed to are standard accuracy, F1, recall and precision. For the latter 3, the macro average of each metric was used unless otherwise stated. Since this weights each class equally, it is a better indicator of performance on an imbalanced dataset, where a micro or weighted average might obscure a model's failure to on the minority classes.[16]

3.1 ANN Methods

Recall at 12 different values of k and runtimes for the initial index setup and the retrieval of the same K-nearest-neighbours were calculated for all indexing methods as described in 2.2. Ground truth for neighbours was established by cosine similarity, as this is a distance metric that better captures the similarity between high-dimensional vectors than Euclidean distance, for example, which suffers from the curse of dimensionality (Bellman et al., 1957). Figures 2 and 3 depict the recall and runtime outcomes respectively.

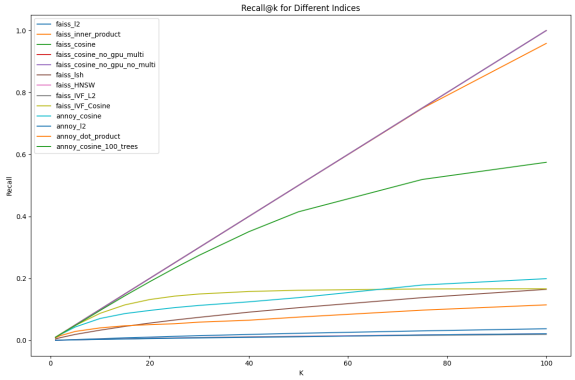


Figure 2: Recall@K for Different ANN Methods

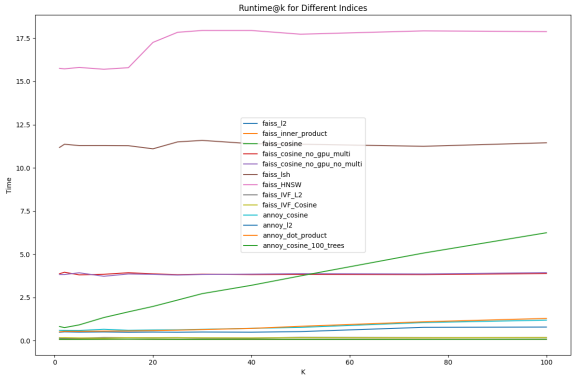


Figure 3: Runtimes for Different ANN Methods

All FAISS flat indices that used cosine similarity as the distance metric provided perfectly accurate recalls, as expected. The ANNOY indices, even when defined on cosine similarity as the metric, did not perform as well, although the increase in number of trees did provide significant improvements. The performance of the non-flat indices was very poor, both in terms of runtimes and recall, and since the gpu-implementation of FAISS was exceptionally fast, even

[14] Query used and teh response obtained can be found in the Appendix.

[15] Technically, this contravenes the Active Learning 'framework' as the pool is theoretically unlabelled, and looking at the labels would be 'cheating'. Had this approach been advanced, it would have resulted in the method being considered quasi-AL.

[16] The trade-off of this is that the metrics sometimes appear frighteningly low - the goal here was not to build an exceptional classifier, but rather to assess different methods.

for flat indices, there would be no reason to make use of a less accurate, approximate index. Therefore a flat cosine similarity index was chosen for integration into the AL pipeline for clustering calculations. The gpu-accelerated flat index was quicker than both the multithreaded and standard CPU implementations, which did not differ between themselves significantly. Nonetheless, experiments for distance-based sampling methods showed an improved performance in the AL pipeline when using an LSH index as opposed to a flat cosine index, where the former obtained an F1 score of 0.25 as opposed to the latter's 0.18 and a random sampling technique's 0.24. Thus, the cluster methods and distance methods ultimately made use of different indices.

3.2 Active Learning Outcomes

As detailed in 2.3, an independent assessment of the clustering without the AL framework, can be found in the Appendix. However, the results were not entirely satisfactory, with a held-out test set accuracy of 0.65 and a recall of 0.14. Nonetheless, it was still integrated as an AL selection strategy, in the understanding that it would aid the sampling process, in a similar manner to how the LSH index objectively performed worse, but integrated better into the AL pipeline (see 3.1 above). Neither of the non-random centroid initialisation nor the dynamic clustering methods had a significant impact on the quality of the clusters or the outcomes on the Active Learning method.
Individually, both non-random initialisation techniques for the initial train set in the active learning pipeline proved insignificant. A combination of both methods, however, resulted in a minor improvement on a random initialisation technique using a randomly sampled AL pipeline, but proved worse when the sampling method was distance-based. Table 1 summarizes these results. The non-random initialisations were not assessed in combination with a clustering-based selection.

Init. Method	Metric	Random Sampling	Distance Sampling
Random	$F1$	0.24	0.25
	Recall	0.22	0.22
Class Coverage	$F1$	0.24	0.18
	Recall	0.22	0.20
Cluster Coverage	$F1$	0.24	0.22
	Recall	0.21	0.20
Combined Coverage	$F1$	0.23	0.25
	Recall	0.22	0.23

Table 1: Outcomes of Train-Set Initialisation Methods for Different Sampling Techniques

Additionally, Table 1 shows how the distance-based sampling technique[17] did not hold up individually, and was outperformed by a simple random selection for any initialisation method. It did, however, integrate well when used in conjunction with a clustering-based method. An initial experiment was conducted using 5 sampling techniques - the cluster and distance-based[18] techniques introduced here, alongside random, margin and entropy sampling, to be used as standard Active Learning Baselines. Figure 4 depicts these outcomes. A second set of experiments were run, this time comparing random, cluster-based and distance-based sampling methods, alongside a combined measure, which used half of the sampling budget on the cluster method and the other half on the distance method. These results can be found in Figure 5. All experiments followed the parameters detailed in 2.3.4, and all methods were assessed for accuracy, F1, precision and recall. Additionally, each set of experiments controlled the train and test sets, ensuring a fair comparison for all methods. Finally, the same 2 experiments were run on each of the 4 evaluation datasets. Table 2 displays the most significant results[19], while all plots for each

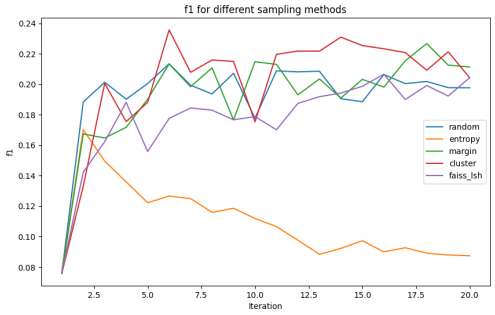


Figure 4: F1 Score for Different AL Sampling Techniques

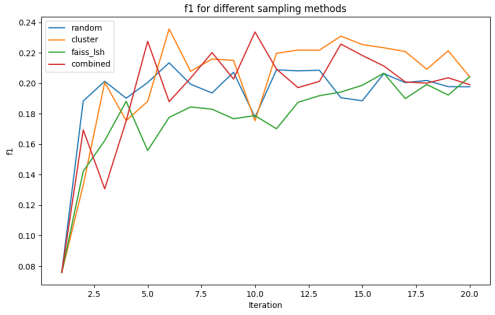


Figure 5: F1 Score for Different AL Sampling Techniques, including the Combined Approach

[17] The average distance method (see 2.3.1) proved the most robust, and is the only one for which results are reported.

[18] Note that in the figures depicting these results, the distance method is labelled as *faiss_lsh*, owing to the use of the index for distance-calculations

[19] Precision and accuracy results have been omitted, as well as the entropy-based sampling, for brevity reasons. Additionally, the Legal Citations & Product Reviews datasets are not displayed, but are included when calculating the average score. All results can be found in the plots in the Appendix

metric, experiment and dataset can be found in the Appendix.

Sampling Method	Metric	IMDb	Papers	E-Co-mmerce	Mean
Random	$F1$	0.196	0.732	0.943	0.5538
	Recall	0.177	0.683	0.941	0.546
Margin	$F1$	0.213	0.788	0.955	0.5748
	Recall	0.178	0.75	0.951	0.566
Clustering	$F1$	0.201	0.752	0.951	0.5636
	Recall	0.177	0.702	0.952	0.555
Distance	$F1$	0.201	0.751	0.940	0.5584
	Recall	0.178	0.702	0.938	0.548
Combined	$F1$	0.198	0.753	0.949	0.5614
	Recall	0.176	0.702	0.947	0.553

Table 2: Outcomes of Train-Set Initialisation Methods for Different Sampling Techniques after 20 Iterations

As can be seen, the clustering method obtained better results than a random sampling method, and than distance sampling. Additionally, the integration of the distance and clustering methods produced slightly worse results than pure clustering. Margin-based sampling, however, consistently beat out all other methods, while entropy-based sampling consistently lost.

4 Discussion

Overall, the distance based sampling method was less effective than the clustering method, yet both slightly outperformed a random sampling, showing promise and potential for further study of these methods. Margin sampling, however, consistently outperformed all other methods, highlighting the robustness of uncertainty-based approaches in active learning. Another key finding is the performance of the LSH index within the AL pipeline, despite its sub-optimal recall in standalone ANN evaluations. This suggests that a more diverse neighborhood of samples, even if not strictly accurate, can improve the learning process in imbalanced settings.

The study also faced several limitations. First, the IMDb dataset, on which the research was conducted, may have constrained the evaluation due to the generality of the textual descriptions and severe class imbalance. This lack of specificity, in turn, made it very difficult to separate the data according to class, and as a result, the clustering algorithm struggled to effectively separate the data into its true classes. This is also indicated by the LSH index performing better within the pipeline despite it being worse in terms of pure neighbour recall - the neighbouring vectors do not necessarily belong to the same class. Second, although the models used to embed the vectors were chosen for their general purpose embeddings, a higher dimensional embedding that can better capture the full semantic relationships between sentences might have been more informative for this task, especially considering the lack of other features. Third, computational resource limitations restricted the use of larger datasets and higher-dimension embeddings - in which case the ANN-based methods may have outperformed traditional margin sampling as dimension increases.

Future work should focus on exploring higher-dimensional embeddings and larger datasets to better assess the impact of ANN-based methods in active learning. Additionally, integrating traditional sampling techniques within the proposed ANN methods, such as within-cluster margin sampling (as opposed to the Least Confident Uncertainty used here) or using more advanced selection tecniques within each cluster, as well as further experimentation with different ANN indices or hybrid models using varying split ratios (as opposed to the even split here), could further enhance performance. Expanding the research to include hierarchical or multi-label classifications could also provide insights into the scalability and adaptability of these methods in more complex settings.

Acknowledgements

We would like to thank the users who uploaded their datasets to Kaggle for research use, as well as Meta and Spotify for making their FAISS and ANNOY libraries public access. Finally, we would like to thank the course staff for a meaningful and informative course concluding in a challenging project from which we gained immensely.

References

Aggarwal, U., Popescu, A., & Hudelot, C. (2021, January). Minority class oriented active learning for imbalanced datasets. In *2020 25th international conference on pattern recognition (icpr)*. IEEE. Retrieved from http://dx.doi.org/10.1109/ICPR48806.2021.9412182 DOI: 10.1109/icpr48806.2021.9412182

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3), 176–204.

Bellman, R., Bellman, R., & Corporation, R. (1957). *Dynamic programming*. Princeton University Press. Retrieved from https://books.google.co.il/books?id=rZW4ugAACAAJ

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. Retrieved from http://dx.doi.org/10.1613/jair.953 DOI: 10.1613/jair.953

Coleman, C., Chou, E., Katz-Samuels, J., Culatana, S., Bailis, P., Berg, A. C., ... Yalniz, I. Z. (2021). *Similarity search for efficient active learning and search of rare concepts*. Retrieved from https://arxiv.org/abs/2007.00077

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (p. 127–136). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1321440.1321461 DOI: 10.1145/1321440.1321461

Genossar, B., Gal, A., & Shraga, R. (2023, December). The battleship approach to the low resource entity matching problem. *Proceedings of the ACM on Management of Data*, *1*(4), 1–25. Retrieved from http://dx.doi.org/10.1145/3626711 DOI: 10.1145/3626711

Hu, R., Delany, S. J., & MacNamee, B. (2009). *Sampling with confidence: Using k-nn confidence measures in active learning.* Retrieved from https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1050&context=scschcomcon DOI: 10.21427/D7H90Z

Joshi, A., Porikli, F., & Papanikolopoulos, N. (2012). Coverage optimized active learning for k - nn classifiers. In *2012 ieee international conference on robotics and automation, icra 2012* (pp. 5353–5358). Institute of Electrical and Electronics Engineers Inc. (2012 IEEE International Conference on Robotics and Automation, ICRA 2012 ; Conference date: 14-05-2012 Through 18-05-2012) DOI: 10.1109/ICRA.2012.6225054

Kashnitsky, Y. (2020). *Hierarchical text classification.* Kaggle. Retrieved from https://www.kaggle.com/dsv/1054619 DOI: 10.34740/KAGGLE/DSV/1054619

Rittler, N., & Chaudhuri, K. (2023). *A two-stage active learning algorithm for k-nearest neighbors.* Retrieved from https://arxiv.org/abs/2211.10773

Appendix

Code

All code, datasets used and results can be found in the associated GitHub Repository.

Evaluation Dataset Label Distributions

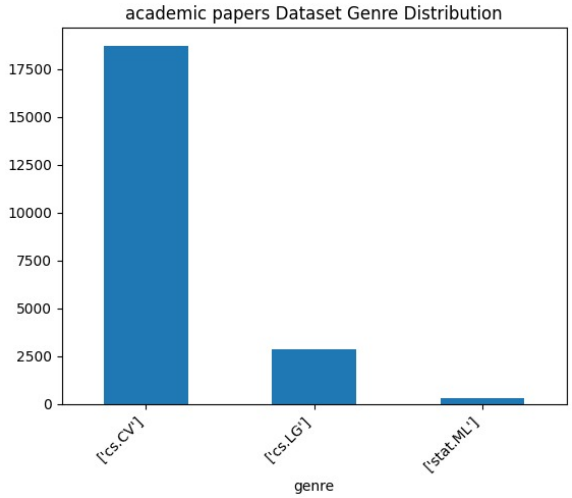


Figure 6: Class Distribution for the arXiv Papers Dataset

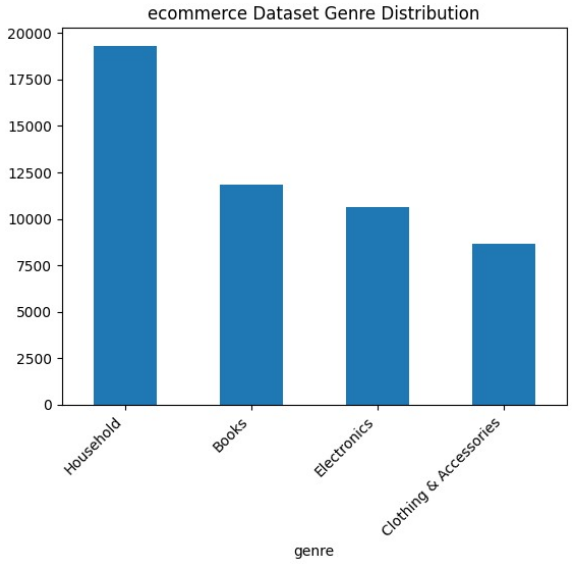


Figure 7: Class Distribution for the E-Commerce Dataset

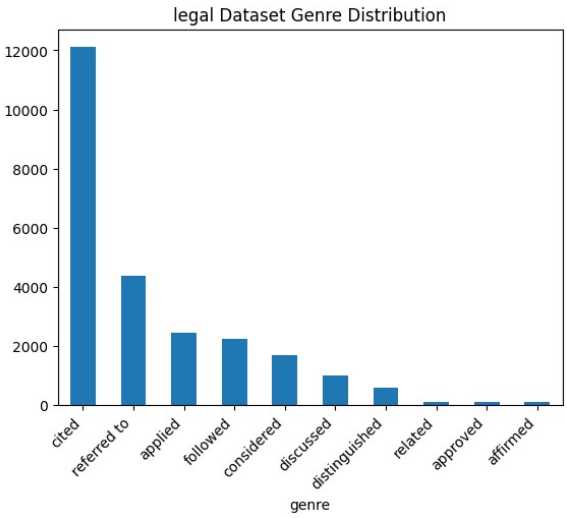


Figure 8: Class Distribution for the Legal Citations Dataset

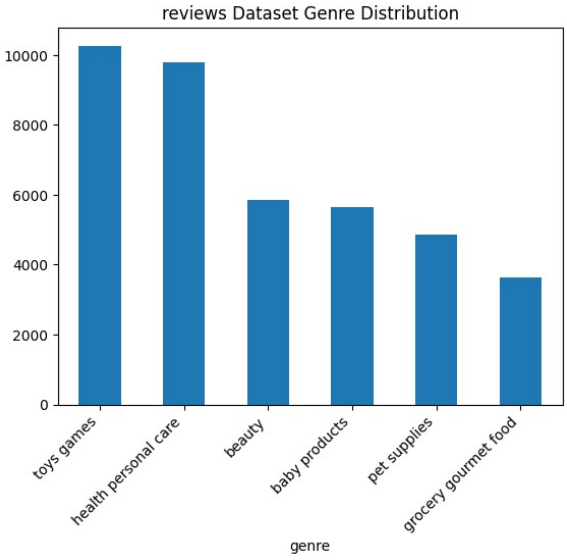


Figure 9: Class Distribution for the Product Reviews Dataset

Query for LLM-based Genre Descriptions as Initial Centroids

The query used is given as follows: "I am wanting to train a language model to classify films according to genre. The following is a list of genres: Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, Western. For each genre, please write a short description of the genre that I can embed for training. Please make sure to exclude any stopwords or punctuation marks, and make sure the answers are in a list format that I can copy into my code."
The obtained response was: genre_descriptions = ["high energy scenes fast paced action sequences physical conflict", "exciting journeys exploration daring quests often unknown worlds", "animated visuals storytelling non-realistic characters fantasy elements", "life stories famous people significant real events personal growth", "humor lighthearted situations witty dialogue fun storytelling", "illegal activities investigations underworld organized crime criminals", "intense character development emotional narratives often personal struggles", "suitable all ages positive themes focus family relationships children", "supernatural magical worlds imaginary creatures often beyond normal reality", "depictions significant past events important historical periods real figures", "designed create fear suspense often include supernatural or violent elements", "centered around music performance songs tell story advance plot", "focus music-based storytelling often include dance performance large-scale productions", "whodunit tension investigation unraveling mysteries complex plots", "focus love relationships emotional connections romance personal struggles", "futuristic science technology space exploration advanced concepts often speculative", "focus sports competition physical challenges athletes victory defeat", "intense suspense heightened tension often dangerous situations psychological drama", "focused on military warfare combat battle strategies soldiers conflict", "stories western frontier rural settings cowboys lawmen outlaws"]

Evaluation of Clustering Method

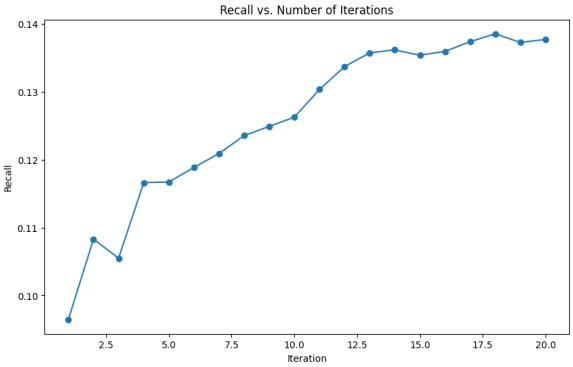


Figure 10: Recall over time for FAISS-based ANN clustering

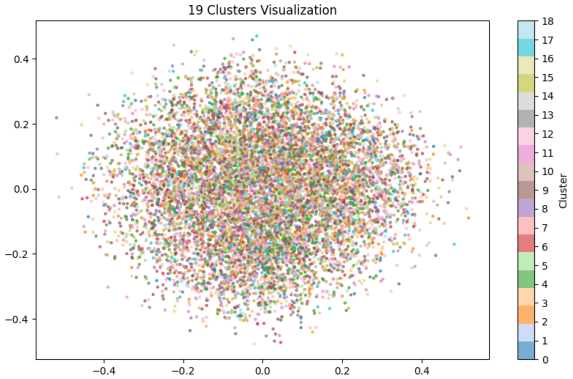


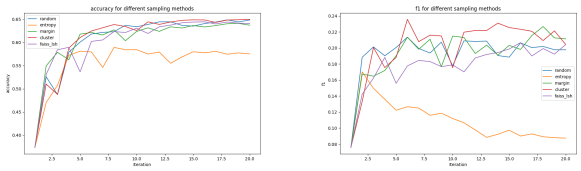
Figure 11: 2-dimensional Visualisation of Clustering Result

Further Plots of Results for All Datasets

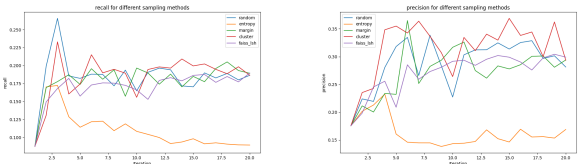
All datasets contain plots of Accuracy, F1, Precision & Recall. For the IMDb dataset, the F1 plots appear in the main text and are omitted here.

IMDb Dataset

Note that the F1 plots for each experiment appear in the text's main body.

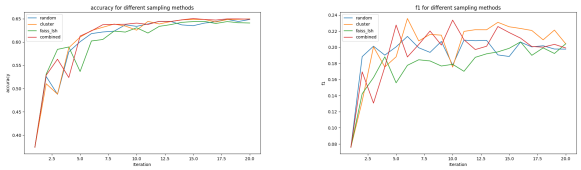


(a) Accuracy (b) F1

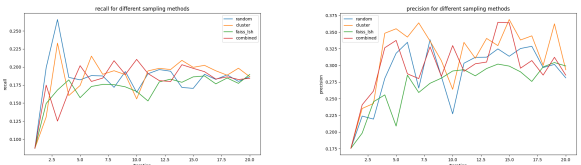


(c) Recall (d) Precision

Figure 12: IMDb Dataset Evaluation Metrics for Different Methods; Experiment 1



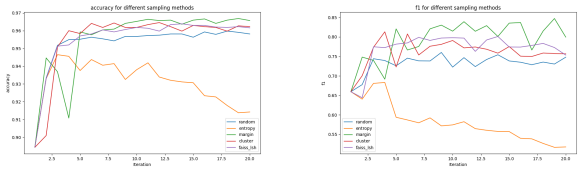
(a) Accuracy (b) F1



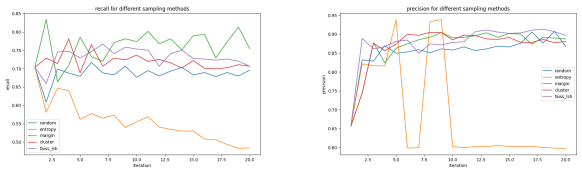
(c) Recall (d) Precision

Figure 13: IMDb Dataset Evaluation Metrics for Different Methods; Experiment 2

Academic Papers Dataset

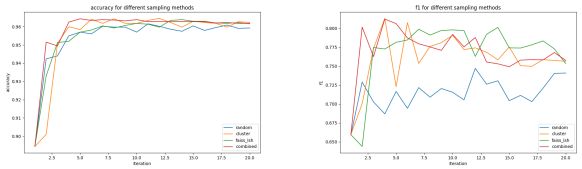


(a) Accuracy (b) F1

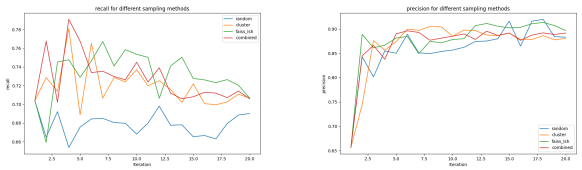


(c) Recall (d) Precision

Figure 14: arXiv Papers Dataset Evaluation Metrics for Different Methods; Experiment 1



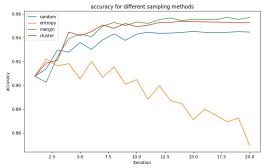
(a) Accuracy (b) F1



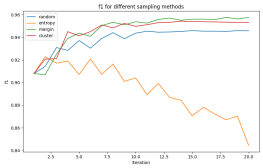
(c) Recall (d) Precision

Figure 15: arXiv Papers Dataset Evaluation Metrics for Different Methods; Experiment 2

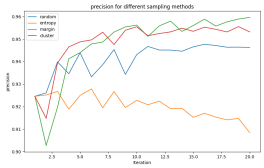
E-Commerce Dataset



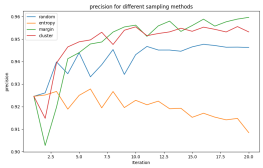
(a) Accuracy



(b) F1

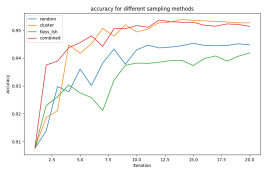


(c) Recall

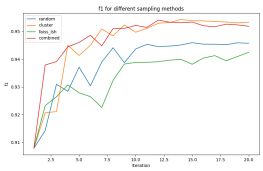


(d) Precision

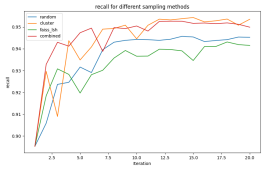
Figure 16: E-Commerce Dataset Evaluation Metrics for Different Methods; Experiment 1



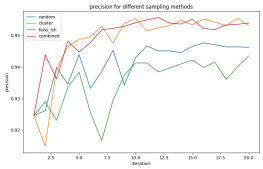
(a) Accuracy



(b) F1



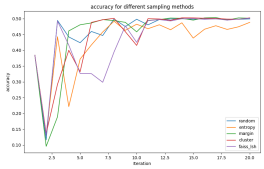
(c) Recall



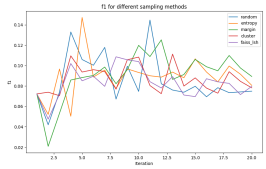
(d) Precision

Figure 17: E-Commerce Dataset Evaluation Metrics for Different Methods; Experiment 2

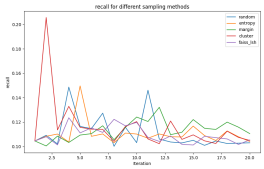
Legal Citations Dataset



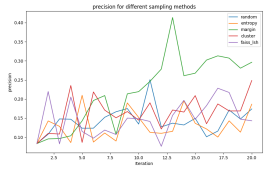
(a) Accuracy



(b) F1

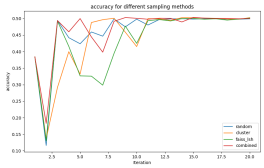


(c) Recall

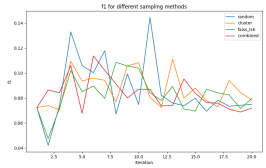


(d) Precision

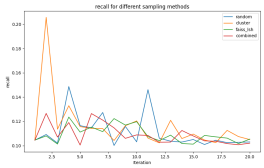
Figure 18: Legal Citations Dataset Evaluation Metrics for Different Methods; Experiment 1



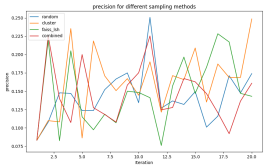
(a) Accuracy



(b) F1



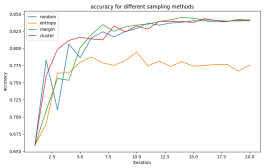
(c) Recall



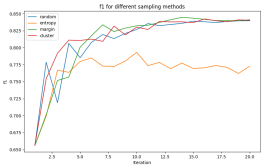
(d) Precision

Figure 19: Legal Citations Dataset Evaluation Metrics for Different Methods; Experiment 2

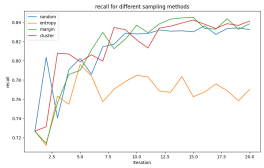
Product Reviews Dataset



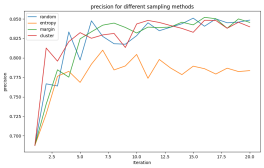
(a) Accuracy



(b) F1

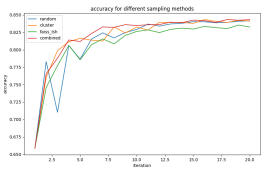


(c) Recall

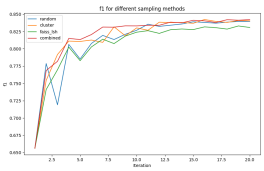


(d) Precision

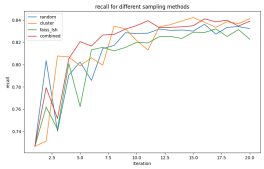
Figure 20: Product Reviews Dataset Evaluation Metrics for Different Methods; Experiment 1



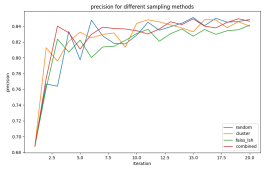
(a) Accuracy



(b) F1



(c) Recall



(d) Precision

Figure 21: Product Reviews Dataset Evaluation Metrics for Different Methods; Experiment 2