An Empirical Study on Active Learning for Multi-label Text Classification

Mengqi Wang
wangmengq@deakin.edu.au

Ming Liu
m.liu@deakin.edu.au

Abstract

Active learning has been widely used in the task of text classification for its ability to select the most valuable samples to annotate while improving the model performance. However, the efficiency of active learning in multi-label text classification tasks has been under-explored due to the label imbalanceness problem. In this paper, we conduct an empirical study of active learning on multi-label text classification and evaluate the efficiency of five active learning strategies on six multi-label text classification tasks. The experiments show that some strategies in the single-label setting especially in imbalanced datasets.

1 Introduction

Active Learning (AL) has been applied in many Natural Language Processing (NLP) tasks due to its efficiency in improving model performance with limited annotation cost. Most works in AL have focused on developing strategies for single-label text classification (Tong and Koller, 2001; Hoi et al., 2006), Named Entity Recognition (Tomanek and Hahn, 2009; Shen et al., 2004, 2017) and Neural Machine Translation (Zhang et al., 2018; Peris and Casacuberta, 2018; Zhao et al., 2020). More recently, multi-label text classification (Liu et al., 2017; Pant et al., 2019; Liu et al., 2021) has received considerable attention since many text classification tasks are multi-labeled, i.e., each document can belong to more than one category. Take news classification as an example, a news article talking about the effect of the Olympic games on the tourism industry might belong to the following topic categories: *sports*, *economy* and *travel*. The challenge of multi-label text classification lies in three aspects: (i) heavily imbalanced labels, i.e. only a small amount of labels have high frequency while others exhibit extremely low frequency; (ii) sparse label correlation, where some labels may be correlated with others, but the correlation is weak; and (iii) hierarchical label structures, this is prevalent in many scientific document indexing, e.g. arXiv or PubMed (Lu, 2011).

Given the above challenges, we raise the research questions: Are the commonly used strategies in single-label text classification still applicable for the multi-label setting? Will they always benefit classification performances? To answer these questions, We conducted an empirical study to evaluate the effectiveness of five AL strategies on six prevalent multi-label text classification datasets. Our experiments show that the strategies commonly used in single-label text classification can have some effectiveness under multi-label settings. However, their performance is not consistent and highly dependent on the label distribution of the datasets. The main findings of our work are as follows:

- **The common AL strategies used in the single label classification are not robust for all multi-label setting.**
- **Diversity strategies consistently outperform other strategies across different dataset sizes and models.**
- **Larger and imbalanced dataset will heavily degrade the performance of common active learning strategies**

2 Active Learning on Multi-label Text Classification

We consider multiple widely-used AL strategies to investigate their different performance on multi-label text classification, including **Least Confidence (LC)** (Culotta and McCallum, 2005), **KMeans** (Kang et al., 2004), **Max Entropy** (Lewis and Gale, 1994), **Deep Bayesian Active Learning(BALD)** (Houlsby et al., 2011), **Monte Carlo (MC) Dropout** (Gal et al., 2017) and **Coreset** (Geifman and El-Yaniv, 2017; Sener and Savarese,

The Fourth Workshop on Insights from Negative Results in NLP, pages 94–102

Algorithm 1 Pool-based multi-label active learning

Input: Initial labeled set L, unlabeled set U, query budget $\mathcal{B}$, model parameter Θ, annotation cost per round b, query strategy Q

Output: The final classifier $\hat{\Theta}$

Initialize Θ_0 with L
for $t \in 1, \ldots,$ **do**
  $\{(\mathrm{x}_i, y_i)_{i=1}^{b}\}^t \leftarrow \text{Query}(U, Q, \Theta_{t-1})$
    ▷ Use strategy Q to select b examples
  $L \leftarrow L + \{(x_i, y_i)_{i=1}^{b}\}^t$
  $U \leftarrow U - \{(x_i, y_i)_{i=1}^{b}\}^t$
  $\Theta_t \leftarrow \text{retrainModel}(\Theta_{t-1}, L)$
  if $b * t > B$ **then** ▷ If budget exhausted
    $\hat{\Theta} \leftarrow \Theta_t$; **break**
return $\hat{\Theta}$

2018). **Random Sampling**, also known as passive learning, randomly selects instances for annotation and serves as a baseline for comparison with other AL strategies. **LC** is one of the most common approach to select queries in active learning, in which it uses the probability to measure how uncertain the model is towards the instances. **KMeans** clustering unlabeled data samples based on their feature representations, and then selecting the samples closest to the cluster centres for labeling. This strategy can help improve the efficiency and effectiveness of the active learning process by focusing on the most representative samples in each cluster. **Max Entropy** measures the confidence of the model using entropy (Shannon, 2001). It ranks all instances in U by the posterior class entropy under the model $H_\theta = -\sum P_\theta(Y \mid X) \log P_\theta(Y \mid X)$, and selects the top unlabelled instances to be labelled by the expert. **BALD** (Houlsby et al., 2011) is another commonly used uncertainty-based AL strategy, which maximizes the mutual information between the predictions and model posterior to achieve maximum information gain. **MC Dropout** selects samples based on their representativeness. As its name, it uses the MC dropout on inference circles, where the uncertainty is measured by the fraction of models across MC samples that disagree with the most popular choice (Siddhant and Lipton, 2018). **Coreset** (Geifman and El-Yaniv, 2017; Sener and Savarese, 2018), is one of the most popular diversity-based querying criteria, which selects the best representation of the dataset using the farthest-first traversal algorithm.

Algorithm 1 shows the pseudo code of our AL loop, given a fixed budget and an initial labeled set L, we try each strategy for the multi-label text classification tasks. In each AL iteration, we acquire b labelled examples, this process is repeated until the budget is exhausted.

3 Experiments

Datasets

Table 1 shows the statistics of the benchmarking datasets that used in the experiments. The datasets vary in size and cover both news and scientific documentation. We took the summary textual context and the corresponding labels for each data set to be the final classification target. All data sets are long-tailed distributed, i.e., only a small portion of labels frequently appear, majority of the label rarely appears in the data. **Web of Science (WOS)** (Kowsari et al., 2017), contains 46,985 documents with 134 categories includes 7 parents categories. All the documents are the published papers from the Web of Science[1] which is a publisher-independent global citation database. All three versions of WOS have been used in this work: WOS-46985, WOS-11967 and WOS-5736. **Arxiv Academic Paper Dataset (AAPD)**[2] (Yang et al., 2018) consists of 55,840 papers abstracts from arXiv[3] in the field of computer science, along with their corresponding subjects. Each paper may have multiple subjects, with a total of 54 subjects included in the dataset. The objective is to predict the appropriate subjects for an academic paper based on the content of its abstract. **Reuters-21578**[4] (Thoma, 2017), is a collection 10,369 news articles appeared on Reuters newswire in 1987. **Yelp Review**[5] is a modified version of the Yelp reviews dataset, consisting of reviews extracted from the Yelp Dataset Challenge 2017. In this dataset, the business label and rating label together are considered as the multi-label for each review.

AL Process

As shown in Figure. 1, we randomly select a tiny portion of initialized data from each dataset to warm-start the classification model. The portion

[1] https://www.webofknowledge.com/
[2] https://github.com/lancopku/SGM
[3] https://arxiv.org/
[4] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[5] https://github.com/rnyati/Yelp-Dataset-Classification-

Dataset	*Size*	*Initial*	*Labels*
WOS5736	5736	1%	11
WOS11967	11967	1%	35
WOS46985	46985	5%	134
AAPD	54840	1%	54
Reuters-21578	10788	1%	168
Yelp Review	208869	5%	466

Table 1: Multi-label text classification datasets.

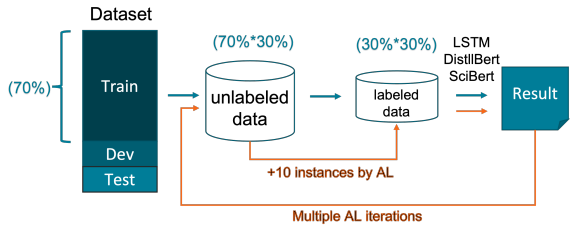


Figure 1: The AL Process in the experiment. Every dataset is divided into three subsets: train, dev, and test. The training data comprises approximately 70% of all samples in the dataset, while the remaining 30% is deemed as unlabeled data. AL strategies are employed to select a few (i.e. 10) instances from the unlabeled data pool, and their labels are then used as the results of the 'human-annotation' process. Multiple rounds of selection are performed until the budget is exhausted.

of initialized random samples ranges between 1%-5% of the 70% training data (see Table 1). We then take the remaining 69% of training data as the unlabeled data pool, which different active learning strategies can actively query. Considering the varying sizes of different datasets, we choose different sizes of annotation budgets, which represents the total number of instances we queried from the selection process. The instances in the unused budget pool will be randomly divided into equal-sized batches to ensure comparable results. The number of selected samples is equally split during each iteration. For each iteration, a batch of samples was identified, and the model was retrained for 20 epochs. The batch size for each dataset is set to 50 follows (Gui et al., 2021). The active querying process stops when all budgets of queried instances are used. Therefore, the batch size setting for different active strategies will be the same for each dataset. We run each strategy on all six datasets 10 times and report the average as the experiment results for evaluation.

Experiment Setup

We conduct the experiment in batch mode, following the traditional pool-based AL scenario(Settles, 2009). To include the popular Bert-based model in our comparison, we adapt the AL strategies following (Ein-Dor et al., 2020). We use **LSTM**(Hochreiter and Schmidhuber, 1997), **DistilBert**(Sanh et al., 2019) and **SciBert**(Beltagy et al., 2019) models. The experiment was implemented by modifying the previous work of large-scale multi-label text classification[6] and incorporating AL settings.

Evaluation Metrics

We use the most representative evaluation metrics for multi-label text classification: Micro-F1 (Huang and Zhou, 2013; Gao et al., 2016; Yu et al., 2020). Micro-F1 score is also known as the micro-averaging of F1 score or simply 'the accuracy' of the multi-label classification problems. It measures the proportion of correctly classified data samples out of all data. As the Micro-F1 score increases, the performance of multi-label text classification improves.

Results

We present the results for all mentioned AL strategies in Section 2. Figure 2 , Figure 3 and Figure 4 show the performance of all strategies on different datasets. We observed that only part of AL strategies improve the accuracy of multi-label text classification among different datasets. The only very promising dataset is Reuters, where all AL strategies outperformed the random baseline on all three models. In most datasets, the random baseline was outperformed by other strategies, even when the baseline performs well, such as in WOS5736.

From a model perspective, AL strategies adapted to DistilBert and SciBert are more robust than those adapted to LSTM. With the boost of the two versions of Bert model, AL strategies can be effective on more datasets in both news and scientific domains. However, AL strategies on the LSTM model provide negative results in both domains. This suggests that without suitable pre-trained models, the AL strategies cannot provide promising results. This can be an important insight for future work, as AL's ability to actively query the most informative samples can better leverage pre-trained models.

[6] https://keras.io/examples/nlp/multi_label_classification/

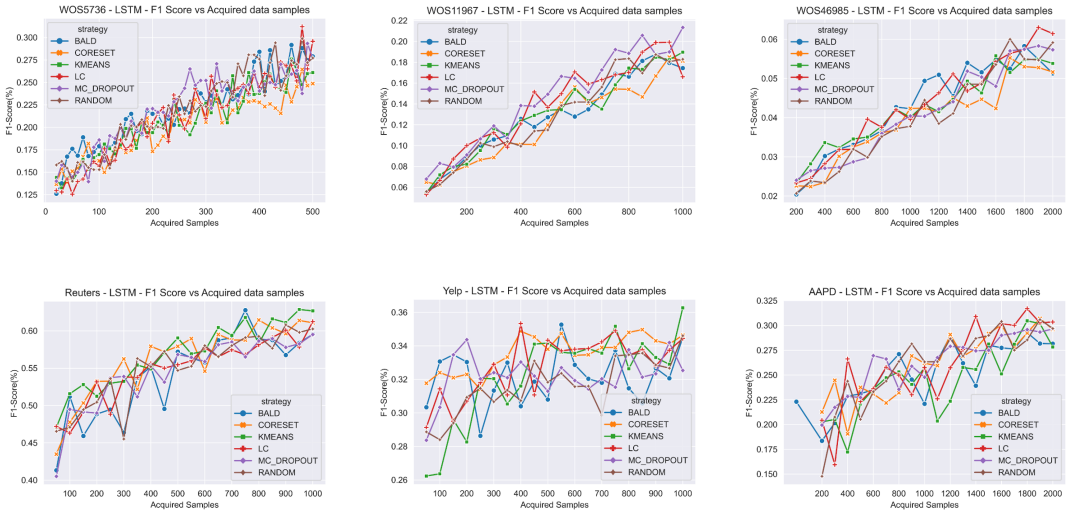


Figure 2: AL Strategies on LSTM

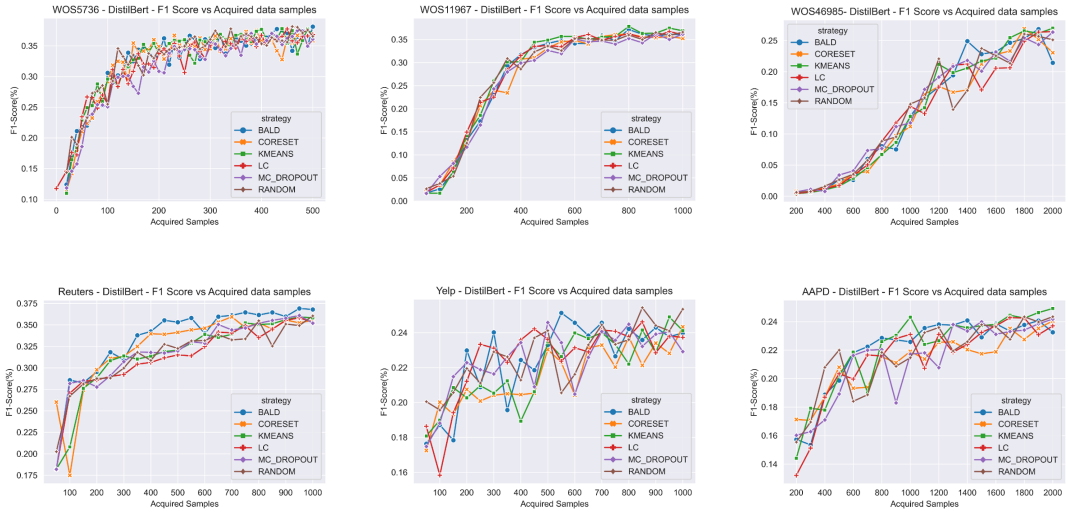


Figure 3: AL Strategies on DistilBert

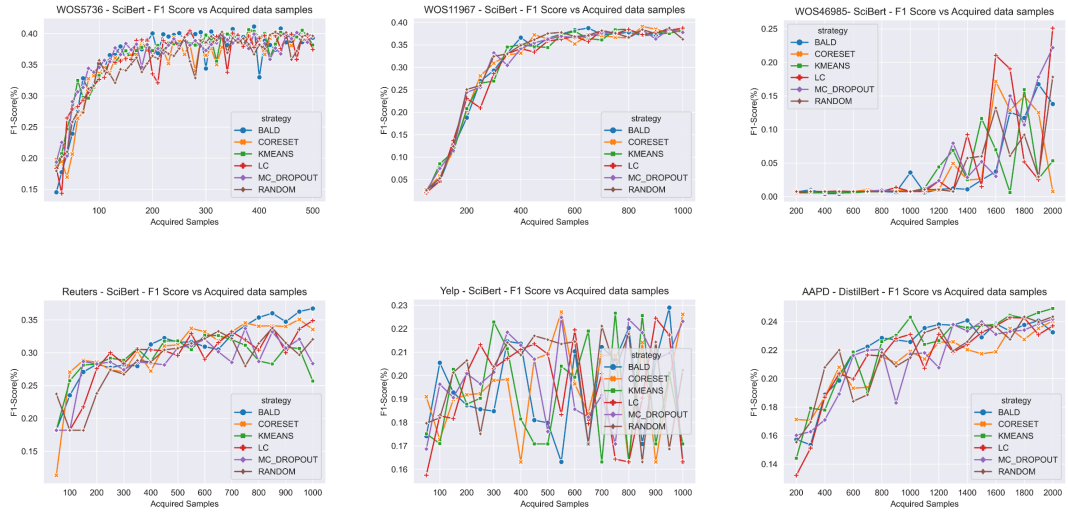


Figure 4: AL Strategies on SciBert

While AL strategies can outperform the Random baseline in multi-text classification using DistilBert and SciBert, there is no single strategy that consistently outperforms all others. For example, BALD in AAPD and Reuters underperforms compared to random. It is natural that no single strategy can outperform all others on all datasets due to the diversity and representativeness of the queried instances, which heavily impacts the effectiveness of AL (Aggarwal et al., 2014; Ren et al., 2021). When the label structure of the original dataset is complex, it is hard for AL strategies to capture both features in the queried instances.The KMEANS strategy achieves the best performance in the larger WOS46985 and AAPD Review datasets. However, in Yelp dataset, it remain comparable to the randombaseline.

Additionally, we do a study to compare the impacts of data sizes on AL performance, the result is presented in the Appendix: Figure 5. We compared the F1-score of three different models, all with the powerful BALD AL strategy, on WOS5736, WOS11978, and WOS46985. In the smaller datasets, WOS5736 and WOS11978, it can be easily observed that BALD effectively improves the F1-score in DistilBert and SciBert after the first ten rounds of actively querying. However, for the larger dataset, WOS46985, BALD only works for DistilBert after ten rounds and takes 20 rounds for SciBert. For all datasets, BALD does not show any effectiveness in all models, as no sudden increase of F1-score can be observed.

We also find that the imbalanced label distribution has an impact on the effectiveness of AL strategies. As shown in Figure 6, the dataset WOS11967, which has the least imbalanced label distribution, has all AL strategies perform better than the other WOS datasets. The accuracy of multi-label text classification with AL improved by over 50% with only one-third of the entire dataset. We plan to conduct a future study to further investigate how label imbalance affects the effectiveness of AL strategies. This research is significant as unbalanced data acquisition can lead to fairness issues that may affect the reliability and validity of machine learning models.

After conducting our initial analysis, we dived deep into the label distribution of the acquired data samples for the WOS dataset in more detail, , the result is presented in the Appendix: Figure 6. We find that the labels in each dataset exhibit an imbalanced distribution, which motivated us to further explore the relationship between AL strategies and the balance of selected data samples in future study. This inquiry is crucial, as unbalanced data acquisition may lead to fairness issues that can significantly affect the validity and reliability of machine learning models.

We also measured and compared the average runtime of one selection iteration for different strategies on all datasets. However, the differences between the runtimes are less than one second. This is understandable, as the different strategies are waiting for the same features from the model's prediction results to decide on the selected samples.

4 Conclusion

In this paper, we explored different Active Learning strategies and its performance on multi-label text classification using a basic neural network model. Our goal is to understand if the popular active learning strategies can prove effective in a multi-label text classification tasks under AL setting. To the best of our knowledge, our work presented the first systematic and comparative study in this context. We observed that unlike single-label text classification, not all strategies can outperform the random baseline. In future work, we plan to perform a deeper analysis of the fairness issue for multi-label text classification under AL setting while exploring more strategies recently published.

References

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. 2014. Active learning: A survey. In *Data Classification*, pages 599–634. Chapman and Hall/CRC.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 7949–7962, Online. Association for Computational Linguistics.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Nengneng Gao, Sheng-Jun Huang, and Songcan Chen. 2016. Multi-label active learning by model guided distribution matching. *Frontiers of Computer Science*, 10(5):845–855.

Yonatan Geifman and Ran El-Yaniv. 2017. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*.

Xiaoqiang Gui, Xudong Lu, and Guoxian Yu. 2021. Cost-effective batch-mode multi-label active learning. *Neurocomputing*, 463:355–367.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Steven CH Hoi, Rong Jin, and Michael R Lyu. 2006. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Sheng-Jun Huang and Zhi-Hua Zhou. 2013. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th international conference on data mining*, pages 1079–1084. IEEE.

Jaeho Kang, Kwang Ryel Ryu, and Hyuk-Chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8*, pages 384–388. Springer.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. 2021. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. *arXiv preprint arXiv:2104.01666*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011.

Pooja Pant, A Sai Sabitha, Tanupriya Choudhury, and Prince Dhingra. 2019. Multi-label classification trending challenges and approaches. *Emerging Trends in Expert Applications and Security*, pages 433–444.

Álvaro Peris and Francisco Casacuberta. 2018. Active learning for interactive neural machine translation of data streams. *arXiv preprint arXiv:1807.11243*.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey.

Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL-04)*, pages 589–596.

Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Martin Thoma. 2017. The reuters dataset.

Katrin Tomanek and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, pages 105–112.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3915–3926.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xiangliang Zhang. 2020. Cmal: Cost-effective multi-label active learning by querying subexamples. *IEEE Transactions on Knowledge and Data Engineering*.

Pei Zhang, Xueying Xu, and Deyi Xiong. 2018. Active learning for neural machine translation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 153–158. IEEE.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

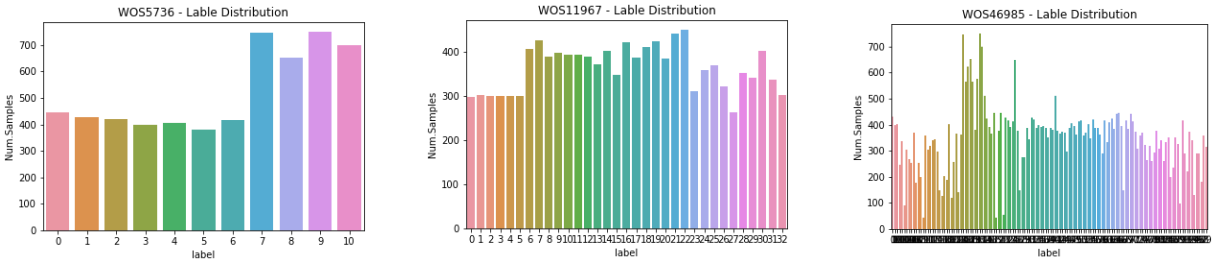
A Appendix

A.1 Data Size.

We performed an exhaustive analysis of the entire dataset by employing three prominent machine learning models, namely Long Short-Term Memory (LSTM), DistilBERT, and SciBERT, in conjunction with three distinct active learning strategies, namely RANDOM, KMeans, and BALD. We systematically augmented the number of acquired samples and meticulously evaluated the resulting changes in F1-score to gain insights into the performance of each model and strategy. This comprehensive evaluation enabled us to identify the most effective combination of model and active learning strategy for optimal performance.

Figure 5: AL strategies on various data sizes and models

A.2 Distribution



((a)) WOS5736 label distribution ((b)) WOS11967 label distribution ((c)) WOS46985 label distribution

Figure 6: Label distribution for three WOS dataset

After conducting our initial analysis, we dived deep into the label distribution of the acquired data samples for the WOS dataset in more detail, as shown in Figure 6. We find that the labels in each dataset exhibit an imbalanced distribution, which motivated us to further explore the relationship between active

learning strategies and the balance of selected data samples in future study. This inquiry is crucial, as unbalanced data acquisition may lead to fairness issues that can significantly affect the validity and reliability of machine learning models.