Enhancing Text Annotation with Few-shot and Active Learning: A Comprehensive Study and Tool Development

Ishika Dhall

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

August 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Ishika Dhall**

Entitled: **Enhancing Text Annotation with Few-shot and Active Learning: A Comprehensive Study and Tool Development**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
Dr. Mirco Ravanelli

______________________________ Examiner
Dr. Mirco Ravanelli

______________________________ Examiner
Dr. Yiming Xiao

______________________________ Supervisor
Dr. Essam Mansour

Approved by ______________________________
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

__________ 2023

____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Enhancing Text Annotation with Few-shot and Active Learning: A Comprehensive Study and Tool Development

Ishika Dhall

The exponential growth of digital communication channels such as social media and messaging platforms has resulted in an unprecedented influx of unstructured text data, thereby underscoring the need for Natural Language Processing (NLP) techniques. NLP-based techniques play a pivotal role in the analysis and comprehension of human language, facilitating the processing of unstructured text data, and allowing tasks like sentiment analysis, entity recognition, and text classification. NLP-driven applications are made possible due to the advancements in deep learning models. However, deep learning models require a large amount of labeled data for training, thereby making labeled data an indispensable component of these models. Retrieving labeled data can be a major challenge as the task of annotating large amounts of data is laborious and error-prone. Often, professional experts are hired for task-specific data annotation, which can be prohibitively expensive and time-consuming. Moreover, the annotation process can be subjective and lead to inconsistencies, resulting in models that are biased and less accurate.

This thesis presents a comprehensive study of Few-shot and active learning strategies, systems that combine the two techniques, and current text annotation tools while proposing a solution that addresses the aforementioned challenges through the integration of these methods. The proposed solution is an efficient text annotation platform that leverages Few-shot and Active Learning techniques. It has the potential to assist the field of text annotation by enabling organizations to process vast amounts of unstructured text data efficiently. Also, this research paves the way for inspiring ideas and promising growth opportunities in the future of this field.

Acknowledgments

I extend my sincere gratitude to Dr. Essam Mansour for his invaluable guidance and unwavering support during my time at Concordia University and the CoDS lab. The research platform provided by the lab was exceptional, and I feel fortunate to have had the opportunity to contribute to top-tier conferences like SIGMOD. Special thanks to Dr. Panos Kalnis, whose brilliance and collaboration on the KGQAn project opened new horizons for universal question-answering on knowledge graphs. Dr. Mansour's mentorship helped me tackle challenges in developing KGQAn, especially in the Question Understanding module. His guidance led to significant progress in obtaining annotated text data and training a massive language model. This experience inspired my current project, an efficient text annotation platform supporting Few-shot learning and active learning.

I owe my deepest gratitude to my rock-solid family, my parents Varinder Dhall and Shashi Dhall, who supported me unconditionally from India. To my brother Garvit Dhall and his partner Shraddha Dhall, thank you for your unwavering support and encouragement. My dear friend Shubham Vashisth, thank you for standing by me through thick and thin, and Reham Omar, your guidance has been invaluable. I am also deeply grateful to my friends and colleagues for their invaluable insights and great ideas, which have paved the way for exciting future endeavors. Last but not the least, I acknowledge the education I received in my home country, India, which laid the foundation for my academic and personal growth.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

With the proliferation of digital communication channels, such as social media and messaging platforms, the volume of unstructured text data has grown significantly in recent years (Messaoudi, Guessoum, & Ben Romdhane, 2022). This has led to an increased demand for natural language processing (NLP) techniques to extract valuable insights and automate various services. NLP enables the analysis and comprehension of human language, making it possible to process unstructured text data effectively. This, in turn, opens up a wide array of applications, including information retrieval, text summarization, sentiment analysis, entity identification, text classification, and translation. One of the significant challenges in the field of NLP-based tasks is the labor-intensive task of manually annotating vast amounts of data. This process, also known as data annotation, is crucial for training machine learning models but can be a time-consuming, costly, and error-prone endeavor (Snow, O'connor, Jurafsky, & Ng, 2008). The increasing volume of unstructured text data exacerbates this challenge, as manual annotation becomes increasingly infeasible. Various techniques have been proposed to mitigate this challenge, such as active learning, weak supervision, and self-supervised learning. Weak supervision relies on noisy labels and heuristics to generate supervision signals (Varma & Ré, 2018) and representations learned using self-supervised learning techniques are significantly inferior to those delivered by fully supervised techniques, so previous work shows that the pretext tasks for self-supervised learning should not be considered in isolation (Zhai, Oliver, Kolesnikov, & Beyer, 2019). On the other hand, active learning techniques allow the model to interact with the annotator and request labels for specific instances, rather than requiring manual annotations for the entire dataset leading to a better model performance with less labeled data therefore, they are more efficient than weak supervision and self-supervised learning. This continues to be a significant area of research in NLP and machine learning. Active learning is a powerful technique for reducing the labor-intensive task of manual annotation in training machine learning models. It utilizes a semi-supervised learning approach, where an iterative process is followed. To further address the challenges faced in active learning and enhance the capabilities of text annotation tools, additional techniques like few-shot learning can be considered. Few-shot learning empowers models to generalize from a limited number of labeled examples, thus boosting adaptability to new and diverse data scenarios (Bennequin, Bouvier, Tami, Toubhans, & Hudelot, 2021). As we delve deeper into the advancements of text annotation tools, it is crucial to consider and evaluate key concepts that contribute to the development of comprehensive and high-performing platforms. Concepts such as flexibility in supporting user-defined models and acquisition techniques, the integration of active learning for efficient annotation, and the exploration of few-shot learning for enhanced model generalization are paramount.

The main aim of this thesis is to explore key concepts for data annotation and conduct a comprehensive study with the evaluation of data annotation tools. By understanding the strengths and limitations of these techniques, we seek to pave the way for more sophisticated and adaptable text annotation platforms, thereby advancing the field of NLP and machine learning. Through this research, we aspire to provide valuable insights and practical guidance for the design and development of cutting-edge text annotation tools. In conclusion, this paper aims to shed light on the key aspects that underpin text annotation tools, offering valuable insights into their advantages and disadvantages. Through an investigation, we also aspire to present a transformative approach to developing a cutting-edge text annotation platform that addresses the challenges and demands of the ever-expanding field of natural language processing.

1.2 Contributions

In this thesis, we aim to present various techniques that can be utilized and combined to form a holistic platform that can address the challenges of data annotation. We present a comprehensive investigation of data annotation tools, analyzing their requirements, drawbacks, and proposing a novel approach to building a holistic and efficient platform. We identify three main research questions that form the backbone of this study.

- This thesis seeks to address essential research questions related to data annotation platforms. The primary objectives are to identify key concepts and characteristics that contribute to the efficacy of such platforms and establish a strong foundation for an advanced annotation tool. By thoroughly understanding these fundamental concepts and features, we aim to develop an annotation tool that effectively caters to the diverse needs of users.

- Another essential research question involves investigating whether the existing systems satisfy the requirements by comparing their features based on the defined concepts. To gain insights into the current landscape of text annotation tools, an in-depth analysis of existing systems is conducted. This examination will provide a comprehensive understanding of the gaps and limitations present in the current state-of-the-art.

- Lastly, the thesis also tries to address the question of how to bring together various latest concepts to create a holistic annotation platform along with an evaluation of techniques used.

1.3 Outline

The paper is organized as follows. Chapter 2 explains the concepts of active learning and various active learning based strategies. Next, the chapter 3 covers few-shot learning concepts and how we can combine active learning and few-shot learning together. Chapter 4 show a comprehensive framework along with a comparison of existing annotation tools. In Chapter 5, we talk about an efficient annotation system outlining the system architecture and implementation details. Also, perform experimentation to evaluate the concept on different tasks. Chapter 6 concludes the paper.

Chapter 2

Efficient data selection - Active Learning

2.1 Overview

Active learning is a machine learning approach that involves an iterative process of selecting and labeling the most informative samples from a large pool of unlabeled data. In active learning, an algorithm actively selects the data instances to be labeled by an oracle (e.g., a human expert or an existing labeled dataset) in order to improve the learning model's performance. The key idea behind active learning is to make the learning process more efficient by selectively querying the labels of data points that are expected to provide the most valuable information to the model. By actively selecting informative samples for labeling, active learning can achieve good performance with fewer labeled examples compared to traditional supervised learning approaches that require labeling the entire dataset. An ideal active learning pipeline would choose the most informative data points using defined heuristics and subsequently pass these data points to an oracle for annotation and inclusion in the final set of training data as shown in Figure 2.1. The figure depicts the iterative Active Learning process, where an initially labeled dataset trains the model. Leveraging active learning techniques and model predictions, the loop intelligently identifies and selects the most valuable samples for labeling, improving the model's performance through iterative sampling and retraining. Active learning optimizes the use of data by recognizing that not all data points contribute equally to the training of the model. It acknowledges that certain data points may be redundant, already well-covered by existing labeled examples, or may provide little additional information for improving the model's performance. By selectively choosing the most informative samples for labeling, active learning focuses on acquiring new information that effectively complements the existing training data, leading to more efficient and effective model training. There can be various active learning scenarios in which the user may ask queries, including Membership Query Synthesis, Stream-Based Selective Sampling, and Pool-Based Sampling, which represent the three primary settings (Settles, 2009). In the membership query synthesis scenario, the learner can request labels for any unlabeled instance in the input space, including queries that the learner newly generated, instead of relying solely on samples from some underlying natural distribution. However, this approach can also pose challenges, particularly in cases involving complex data like natural language. In such scenarios, the learner may have to deal with many gibberish and noisy queries, making it difficult to determine the most informative instances to query. Other active learning scenarios like Stream based query selection can be very useful when an unlabeled instance can be easily sampled from the distribution and the learner can then decide whether the label of that instance is required or not. It is also referred to as sequential active learning as this approach involves sampling instances one at a time from the distribution and allowing the learner to decide whether to query or discard them. To make effective decisions on which instances to query, active learning often relies on informative measures and query strategies. Two very commonly used query strategies are uncertainty-based sampling and query-by-committee. Uncertainty-based sampling selects instances that the model finds difficult to classify. This strategy can be used with various uncertainty measures, including entropy, margin, and least confidence. Alternatively, Query-by-committee relies on multiple models, rather than just one. The idea is to train several models with different parameter settings. Then, when an unlabeled instance is queried, each model predicts its label. The most informative instance is the one that has the greatest disagreement between the models, indicating a high degree of uncertainty (Gilad-Bachrach, Navot, & Tishby, 2005). If there is a small set of labeled data and a large pool of unlabeled data available, queries can be drawn from the pool. This scenario comes under the category of pool-based active learning which has its main application in Text Classification (D. D. Lewis & Gale, 1994; McCallum & Nigam, 1998), Image classification (C. Zhang & Chen, 2002), Speech Recognition (Tür, Hakkani-Tür, & Schapire, 2005), etc. The pool-based approach is different than the stream-based approach as it ranks the entire pool of data before finally querying the instance

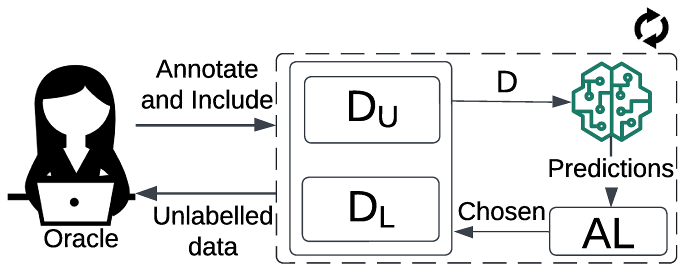

Figure 2.1: Iterative model training using active learning for optimal performance enhancement

while the stream-based approach scans through data sequentially making query decisions one by one.

Active learning has been widely researched and formalized in various works (Ren et al., 2021), with the aim of combining its benefits with the power of deep learning. Although deep learning has made remarkable strides in various fields, it requires a vast amount of labeled data due to its data-hungry nature. Collecting labeled data can be a major challenge, especially when the labeling process requires specialized knowledge of the domain. To maximize the model's performance with a small amount of labeled data, active learning provides an effective solution by selecting the most informative data samples for labeling. However, active learning also has its own challenges, including selecting informative samples for labeling, which can be addressed through query strategy optimization (Ren et al., 2021). It also generally faces the problem of deciding the stopping criteria when the desired performance level has been reached (Vlachos, 2008). Another challenge with active learning is the need for an initial set of labeled data to train an initial set of weights and biases of a model, which are then used for subsequent active learning steps. This issue is commonly experienced in any active learning pipeline and is referred to as the ”cold start” problem. Fortunately, solutions exist to overcome this problem, such as using pre-trained language models for NLP-based problems (Yuan, Lin, & Boyd-Graber, 2020).

In order to investigate if Active Learning can be used to efficiently reduce the cost of sample annotation while retaining Deep Learning's powerful learning capabilities, the research community saw the emergence of DeepAL (Ren et al., 2021). One main challenge faced during combining Deep Learning with Active Learning is limited labeled samples and the high computational cost of training deep learning models from scratch, especially when the task is specific to a certain domain.

Another main challenge in combining the two is dealing with model uncertainty. Many AL techniques like uncertainty-based query strategies rely on the model uncertainty, which is often difficult to quantify accurately. One way that uncertainty manifests in Deep Learning models is through overconfident probability distributions. The final output layers, like softmax, yield probability distributions that are not always calibrated (Guo, Pleiss, Sun, & Weinberger, 2017). This means that the probabilities output by a neural network can't always be trusted to accurately reflect the uncertainty in the model's predictions. To address this issue, post-processing techniques such as calibration and temperature scaling (Rahaman & Thiéry, 2021) (Guo et al., 2017) have been developed to improve the accuracy of model predictions. Deep learning models require a vast amount of labeled data to learn their complex parameters effectively, whereas classical active learning methods rely on a small amount of labeled sample data to learn and update the model. This makes it difficult to combine the two approaches since the small labeled data samples used in active learning are often insufficient to train deep learning models effectively. Moreover, the one-by-one sample query method commonly used in active learning is not suitable for deep learning models due to their complexity and the need for full training until convergence (Zhdanov, 2019a). As a result, alternative batch sampling methods must be employed to select a large subset of data points for labeling, resulting in correlated samples even for moderately small subset sizes. Also, the inconsistency in the processing pipeline is another challenge in combining active learning and deep learning. AL algorithms concentrate on training classifiers based on fixed feature representations, while in deep learning, feature learning and classifier training are jointly optimized. Overall, combining deep learning and active learning is a challenging task that requires careful consideration of the sampling method, labeling budget, and model uncertainty to achieve accurate and scalable results.

2.2 Active Learning Strategies

Active learning strategies have received considerable attention in research and have been extensively studied and documented in the literature, as evident from the surveys conducted by (Ren et al., 2021) and (Settles, 2009). In order to gain a thorough understanding and systematically evaluate the wide range of query strategies and techniques employed in active learning, we will explore various query strategies, with a particular focus on text annotation tasks. By examining these strategies, we aim to identify effective approaches that can enhance the efficiency and effectiveness of active learning in text annotation for a given use case.

2.2.1 Uncertainty and Hybrid Query Strategy

Uncertainty and hybrid query strategy is a popular and straightforward approach due to its simplicity and low computational cost. This approach selects the most uncertain samples in the given distribution to form a batch query set, as these samples are expected to provide the learning model with the most information. This strategy is particularly effective with probabilistic models, and it has been widely used with shallow networks like KNNs and SVMs because these traditional models can accurately obtain the uncertainty of the predictions. A use case of uncertainty-based query strategies can be a simple binary classification task using a probabilistic model. In such cases, the strategy selects the instance whose posterior probability for being positive is closest to 0.5 (D. D. Lewis & Catlett, 1994). For problems with multiple classes, a more specific variant of this technique like Least Confidence (LC) can be used. Here, the instance with the least confident prediction or the class label with the highest posterior probability is queried using equation 1 (Culotta & McCallum, 2005).

$$1 - \max_{y_1,..,y_n} \mathbb{P}[y_1, .., y_n \mid \{\mathrm{x}_{ij}\}] \tag{1}$$

Another example of an uncertainty-based query strategy can be the Margin sampling query strategy (Scheffer, Decomain, & Wrobel, 2001), where the difference between the highest predicted probability and the second-highest predicted probability for a given sample is computed and called the margin. Samples with the lowest margin, i.e., samples with the least certain prediction, are selected for labeling. This simple approach has been proven effective and has been widely used in various applications. For example, (An, Wu, & Han, 2018) claimed that Margin Sampling is the best technique to be used when combining acquisition strategies with deep neural networks like RNNs for text classification. They considered two strategies, namely Query-by-committee (QBC) and LC. They mentioned that QBC is not a very easy technique to be combined with Deep Learning as it can be really computationally expensive and on the other hand, LC may not always choose the most informative instance. Also, although LC was popular in statistical sequence models, it was later found to select longer sequences disproportionately, such as in named entity recognition tasks. To address this issue, Maximum Normalised Log probability (MNLP) (Shen, Yun, Lipton, Kronrod, & Anandkumar, 2017) was proposed, which normalizes the LC equation to mitigate this problem, as shown in equation 2.

$$\max_{y_1,..,y_n} \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}[y_i \mid y_1, .., y_{n-1}, \{\mathrm{x}_{ij}\}] \tag{2}$$

Experiments using this technique showed that models incorporating semantic relationships between words in a sentence outperformed classical methods while using only 25% of the training dataset (Shen et al., 2017). Another system proposed by (Radmard, Fathullah, & Lipani, 2021) generalized the concept of LC and MNLP (Shen et al., 2017) to define a family of acquisition functions that would work well with both long sequences and smaller sub-sequences. Other uncertainty-based query strategies include entropy-based methods, such as Maximum Entropy (Z. Qiu, Miller, & Kesidis, 2017), which generalize well to probabilistic multiclass classifiers. These methods perform comparably depending on the use case when empirically evaluated in different tasks (Settles, 2009) (Settles & Craven, 2008) (Schein & Ungar, 2007). Many DeepAL systems, such as Bayesian dropout approximation (Gal & Ghahramani, 2016a) using model uncertainties, neural link prediction with Activelink (Ostapuk, Yang, & Cudré-Mauroux, 2019), and an approach that uses both feature learning and task learning-based uncertainty (He, Jin, Ding, Yi, & Yan, 2019), follow the concept of uncertainty query strategy. Some other techniques like DeepFool Active Learning (DFAL) (Ducoffe & Precioso, 2018) challenges the efficiency of existing uncertainty-based methods for deep networks. DFAL argues that these techniques are not effective in handling adversarial examples, which are samples that have been slightly modified with specific perturbations to cause the model to misclassify even when it is highly confident in its prediction. To address this challenge, DFAL uses decision boundary-based margin sampling, which has been shown to outperform state-of-the-art methods with faster computation. By leveraging this technique, DFAL provides a more efficient and effective way to handle adversarial examples and improve the performance of deep learning models. Although these active learning techniques can effectively select informative samples for labeling, they may not fully capture the underlying distribution of the data and the relationship between the samples and relevant knowledge about the data distribution (Marbinah, 2021). Consequently, it can lead to the collection of a redundant set of samples at every step. To address this problem, researchers have proposed hybrid query strategies that take into account both the diversity and information volume of samples, particularly in a batch query setting. By incorporating diversity as a criterion for selecting samples, hybrid query strategies can enhance the overall performance of the active learning process. A diverse set of samples ensures that the underlying distribution of the data is fully captured and avoids the collection of redundant samples.

Furthermore, hybrid query strategies can provide more flexibility in the selection of samples, which can improve the efficiency of the active learning process. For example, Exploration-P uses information entropy to evaluate the uncertainty of a sample under a given model and considers the similarity between samples to understand the true data distribution of the feature space (Yin et al., 2017). Similarly, other systems like Diverse Mini-Batch Active Learning (DMBAL) focus on both informativeness and diversity in the mini-batch query setting (Zhdanov, 2019b), while Weighted Incremental Dictionary Learning (WI-DL) incorporates an unsupervised feature learning stage in addition to the supervised fine-tuning stage of deep belief networks (P. Liu, Zhang, & Eom, 2017). Batch Active Learning by Diverse Gradient (BADGE) offers a trade-off between predictive uncertainty and sample diversity without requiring hyperparameter tuning (Ash, Zhang, Krishnamurthy, Langford, & Agarwal, 2020). By utilizing these and other query strategies, deep active learning can achieve high-performance gains with a small set of labeled data. In recent years, several hybrid methods have emerged that combine uncertainty-based and data distribution-based active learning techniques. WAAL (Shui, Zhou, Gagné, & Wang, 2020) is one such method that explicitly incorporates an uncertainty-diversity trade-off in the selection of a query batch. Unlike uncertainty-based methods that may fail to fully utilize the data distribution, or data distribution-based methods that may ignore the structure of the task, WAAL strikes a balance between the two. Another approach, called Variational Adversarial Active Learning (VAAL) (Sinha, Ebrahimi, & Darrell, 2019), utilizes variational autoencoders and a discriminator to learn a latent representation of labeled and unlabeled data points in an adversarial game. VAAL combines uncertainty and representativeness during sampling in a task-agnostic manner. A subsequent work, called TA-VAAL, builds on VAAL and leverages the latent space via ranking conditional generative adversarial networks (RankCGANs) (Saquil, Kim, & Hall, 2018) to possess task awareness (Kim, Park, Kim, & Chun, 2021). These hybrid methods represent an exciting direction for active learning research and may offer promising solutions for tackling the challenges of uncertainty and distribution-based sampling.

2.2.2 Query-by-committee

As mentioned previously, uncertainty sampling is a commonly used active learning strategy that selects informative samples based on the uncertainty estimation of a single model that can bring bias in the system. However, Query by Committee (QBC) overcomes the limitations of uncertainty sampling by involving multiple models or "committee members" in the decision-making process[1]. In QBC, a diverse committee of models is trained using available labeled data. Each committee member independently predicts the labels for unlabeled instances. The committee's disagreement or consensus on these predictions serves as a measure of uncertainty or informativeness. Instances that generate higher disagreement among committee members are considered more uncertain and thus selected for labeling. By incorporating the collective opinion of the committee members, QBC mitigates the bias that may arise from relying on a single learner, which is a limitation of uncertainty sampling. Additionally, QBC facilitates the capture of examples that fall beyond the scope of a single model's perspective. QBC enables the exploration of diverse viewpoints and information sources within the committee, enhancing the robustness and reliability of the active learning process. Leveraging the diversity of the committee members, QBC makes more informed decisions regarding which samples to label. This approach improves the reliability and effectiveness of the process by leveraging the collective decision-making power and diversity within the committee. In query-by-committee, a committee of models C can be represented as:

$$\phi^{(1)}, ..., \phi^{(C)} \tag{3}$$

[1] https://modal-python.readthedocs.io/_/downloads/en/latest/pdf/

for C different hypotheses that align with the labeled set. The objective is to identify the most informative query, which is determined by the instance that elicits the highest disagreement among the committee members in terms of labeling. Various approaches can be employed within the query-by-committee strategy. For instance, query-by-bagging can be used to train a committee of Conditional Random Fields (Seung, Opper, & Sompolinsky, 1992). Vote entropy is another strategy that leverages Kullback-Leibler divergence, an information-theoretic measure quantifying the difference between two probability distributions, to guide the committee's decision-making process (Dagan & Engelson, 1995). A modification of the aforementioned technique was proposed by (Settles & Craven, 2008) specifically for sequence models. This modification involves computing the average Kullback-Leibler divergence scores by summing the marginals at each token position. Similar to vote entropy, the scores are normalized for the length of the sequence. This adaptation allows for the effective utilization of the Kullback-Leibler divergence measure in the context of sequence models, enhancing the accuracy and reliability of the active learning process. Additionally, a variant based on sequence Kullback-Leibler divergence can also be employed in this context called sequence vote entropy (Settles & Craven, 2008). These approaches provide effective means to select informative queries by exploiting the disagreement or entropy measures among committee members. By leveraging diverse hypotheses within the committee, these strategies contribute to the robustness and accuracy of the active learning process.

2.2.3 Density-weighted query strategy

Many active learning heuristics are ineffective when applied to deep learning models in batch settings (T. Wang, Zhao, Lv, Hu, & Sun, 2021). Deep learning architectures such as CNNs, RNNs, etc., require a significant amount of labeled data to learn complex parameters. However, labeling a dataset can be a time-consuming and expensive task. Therefore, the optimal way to choose data points to label given a fixed labeling budget is a crucial question. Classical active learning algorithms choose a single point at each iteration, but for large networks, this is not feasible due to local optimization methods and the need for full training until convergence, making it intractable to query labels one by one. Instead, a large subset must be queried at each iteration, resulting in correlated samples even for moderately small subset sizes. The basic idea of the Density-weighted query strategy is that samples that are more different from labeled samples can bring more information, and samples that are highly representative of unlabeled samples can promote model generalization. It selects a batch of informative samples as representative as possible at each iteration by making use of the geometry of the samples. Active learning methods can be tailored to the batch sampling case by defining the problem as a core-set selection problem, where the goal is to select a small subset of labeled data, called the core set, that is representative of the entire dataset (Sener & Savarese, 2018). This is achieved without using the labels, and the core set is used to train a model that is competitive for the remaining data points. To tackle the unlabeled core-set problem for CNNs, a rigorous bound between the average loss over any subset of the dataset and the remaining data points is provided via the geometry of the data points. The active learning algorithm aims to choose a subset that minimizes this bound. This approach significantly reduces the cost of acquiring labeled data, making it a cost-effective and scalable solution for various machine-learning problems. By employing this approach, deep learning architectures can learn from a smaller amount of labeled data, thus enabling faster and more accurate model training.

Another technique in this category is Discriminative Active Learning (DAL) (Gissin & Shalev-Shwartz, 2019). DAL is a batch-mode active learning algorithm designed for neural networks and large query batch sizes. It leverages the data density information during the active learning process. DAL formulates active learning as a binary classification task, aiming to select examples for labeling in such a way that the labeled set and the unlabeled pool become indistinguishable. One notable advantage of DAL is its ability to sample from the unlabeled dataset in proportion to the data density, without introducing bias towards densely populated regions. It operates by iteratively selecting batches of examples from the unlabeled pool for labeling. It leverages the discriminative power of a trained neural network to identify instances that are most likely to improve the model's performance when labeled. The goal is to create a labeled set that closely resembles the overall distribution and characteristics of the unlabeled pool, thus ensuring that the model generalizes well to unseen data. One key strength of DAL is its capability to sample from the unlabeled dataset in proportion to the data density. This means that the algorithm can effectively prioritize regions of the data space that are sparsely populated, ensuring that important but underrepresented samples are not overlooked during the active learning process. By constructing a core set, the goal is to ensure that the output of an algorithm applied to the core set closely resembles its output when applied to the complete dataset. Building upon the concept of dataset compression through core sets, a new active learning algorithm called Farthest First Active Learning(FF-Active) is introduced in (Geifman & El-Yaniv, 2017). This algorithm introduces a novel approach to querying consecutive points from the unlabeled pool by utilizing farthest-first traversals in the space of neural activation over a representation layer. However, samples in the query batch sample set cannot represent the true data distribution of the feature space due to the insufficient diversity of batch sample sets. To mitigate this limitation, an algorithm called Exploration-P (Yin et al., 2017) employs a deep neural network to learn the feature representation of the samples and subsequently calculate their similarity explicitly. It addresses the challenge of striking a balance between exploitation and exploration by incorporating a balanced approach, considering both the exploitation of learned representations and the exploration of uncertain regions in the early stages of model training. Previous studies (Roy & McCallum, 2001b; X. Zhu, Lafferty, & Ghahramani, 2003) have highlighted a potential drawback of uncertainty sampling and Query by Committee methods, namely their tendency to query outliers. The least specific instance, which typically lies on the classification boundary, may not be representative of the overall data distribution. Consequently, obtaining its label is unlikely to significantly enhance the accuracy of the model on the entire dataset. Both QBC and Expected Gradient Length exhibit similar behavior, as they allocate resources to querying potential outliers based on their controversial nature or the expectation that they would induce substantial changes in the model. While this approach can be valuable in specific scenarios, such as when exploring novel or uncertain regions of the feature space, it may lead to the selection of instances that do not effectively contribute to the overall model performance. To tackle this challenge, a novel active learning method called Information Density is introduced by (Settles & Craven, 2008) where the approach addresses the problem of selecting informative instances for labeling by considering their average similarity to all other sequences in the unlabeled pool, weighted by a parameter that determines the relative importance of the density term. To calculate the information density of an instance, the paper first computes its sequence entropy which measures the base informativeness of the sample. Then, the paper computes the average similarity of the instance to all other sequences in the pool, using a cosine similarity function. One of the potential drawbacks of information density could be the number of required similarity calculations that grow with the number of instances in the pool.

2.2.4 Batch Mode Deep Active Learning

Batch Mode Deep Active Learning (BMDAL) is a query strategy that operates by selecting batches of samples instead of individual ones. This approach is particularly useful for mitigating the inefficiencies of traditional active learning, where the model is trained with only small incremental changes to the training data. Such a training method is prone to overfitting and can be challenging for deep-learning models. BMDAL, on the other hand, allows for the simultaneous querying and training of the model, leading to more efficient and effective learning. To employ BMDAL, an acquisition function is used to score a batch of unlabeled samples from a large pool of data. The top-k samples with the highest scores are then selected in each acquisition step. One popular approach for selecting batches of informative samples is Bayesian Active Learning by Disagreement (BALD) (Houlsby, Huszar, Ghahramani, & Lengyel, 2011), which was applied to various classification and regression tasks in (Gal & Ghahramani, 2016a). In this approach, the mutual information between the model predictions and parameters is used to select the top-k samples with the highest scores. A technique inspired by BALD and extends the work is BatchBALD (Kirsch, van Amersfoort, & Gal, 2019), which jointly selects multiple informative points for deep Bayesian active learning based on a tractable approximation to the mutual information between a batch of points and model parameters. Another such method proposed by (Azimi, Fern, Fern, Borradaile, & Heeringa, 2012) utilizes a Monte Carlo simulation to estimate the distribution of unlabeled examples. These examples are sequentially selected, and the k best samples are chosen for labeling. This approach leverages the benefits of batch sampling to improve the overall performance of the learning process. In (Shui et al., 2020), Wasserstein distance, which considers the joint probability distribution, is employed to unify the querying and training processes in deep batch active learning. By utilizing this distance metric, a comprehensive method is developed that effectively integrates the selection of samples and the model training procedure. Additionally, another approach introduced in (Settles, Craven, & Ray, 2007) is Multiple Instance Active Learning (MIAL). This technique addresses the challenge of learning from labels at mixed levels of granularity. Instead of labeling individual instances, it organizes instances into bags, and the bags are labeled for training. The goal is to learn a function that maps bags to their correct labels while inferring the labels of individual instances within each bag. This method extends the traditional supervised learning strategy and incorporates active query selection strategies inspired by the Multiple Instance setting. Collectively, active learning techniques in batch mode, such as those discussed above, offer efficient and effective ways to train deep learning models. By selecting batches of informative samples, these approaches enhance performance, reduce overfitting, and enable more accurate modeling of the underlying data distributions.

2.2.5 Deep Bayesian Active Learning

Deep Bayesian Active Learning (DBAL) leverages Bayesian deep learning models to enhance classification accuracy by incorporating uncertainty modeling within the network's predictions. This approach has shown promising results in reducing the required training data while improving performance (Gal, Islam, & Ghahramani, 2017). One such notable method was proposed by (Siddhant & Lipton, 2018) applies Bayes-by-Backprop in the context of Deep Active Learning. This method explores various data acquisition strategies across different text classification tasks such as Named Entity Recognition, Sentiment Classification, and Semantic Role Labeling. Their study demonstrates the consistent superiority of the Bayesian approach compared to other methods across all scenarios. Several other systems, including Bayesian Active Learning by Disagreement (Houlsby et al., 2011), BatchBALD (Kirsch et al., 2019), Dropout as a Bayesian Approximation (Gal & Ghahramani, 2016a) and Batch Active Learning via Coordinated Matching (Azimi et al., 2012) operate on similar principles. These systems employ different techniques to incorporate Bayesian concepts and active learning strategies, contributing to the advancement of batch-mode active learning. A reusable library with a systematic study was presented in (Atighehchian, Branchaud-Charron, & Lacoste, 2020) which is a system that demonstrates the effectiveness of partial uncertainty sampling and larger query sizes in improving the efficiency and accuracy of the active learning loop. It is an open-source Bayesian active learning library that promotes further exploration and development in this area. They focus on two techniques that enhance the active learning process. The first technique, partial uncertainty sampling, improves efficiency by selectively sampling instances that exhibit higher uncertainty which accelerates the learning loop by focusing on the most informative data points. The second technique involves increasing the query size, enabling the model to receive larger batches of labeled data at once. This larger query size enhances the speed and efficiency of the active learning process. Bayesian models provide a mathematically rigorous framework for reasoning about uncertainty, although they typically come with high computational costs. Another strategy that casts dropout training in deep neural networks as approximate Bayesian inference in deep Gaussian processes was proposed by (Gal & Ghahramani, 2016b). This theoretical framework enables them to leverage dropout neural networks as a means to model uncertainty, extracting valuable information from existing models that would otherwise be discarded. This addresses the challenge of representing uncertainty in deep learning without compromising computational complexity or test accuracy. Another potential Bayesian Active Learning-based solution can be offered by leveraging uncertainty estimates from Bayesian Neural Networks. However, applying BNNs to large-scale problems requires approximations due to the need for both performance and uncertainty estimation. Therefore, a method that tries to solve the problem of applying Bayesian Neural Networks at a large scale called Deep Probabilistic Ensembles (DPEs) was presented in (Chitta, Alvarez, & Lesnikowski, 2018). DPEs are a scalable technique that approximates deep Bayesian Neural Networks using a regularized ensemble. The key idea is to train multiple models with a novel KL regularization term that encourages diversity among the models, while also approximating the posterior distribution of a BNN. There is an issue of mode collapse in the current state-of-the-art deep Bayesian active learning method, where the acquisition function of the active learning process can become overconfident and focus on a small subset of the data, leading to poor performance. To deal with this problem, the Deep Ensemble Bayesian Active Learning (DEBAL) method was proposed by (Pop & Fulop, 2018) which corrects this deficiency by making use of the expressive power and statistical properties of model ensembles. This allows DEBAL to capture superior data uncertainty and avoid the mode collapse problem, resulting in improved classification performance.

2.2.6 Expected Gradient Length and Model Change

In contrast to existing active learning methods that focus on reducing version space size, some research, such as (Roy & McCallum, 2001a), introduces an active learning approach that directly optimizes expected future error. Gradient-based Active Learning methods such as Expected Gradient Length (EGL) have remained less explored and understood, as highlighted in (Huang et al., 2016). The work in (Huang et al., 2016) specifically opted for EGL due to its unique ability to reduce variance and effectively identify informative samples that differ from those selected based on confidence scores in the context of Speech Recognition tasks. The paper demonstrates the significant benefits of EGL in speech recognition, showcasing its potential to substantially improve performance by reducing word errors and minimizing the reliance on labeled samples. These findings underscore the value of EGL as a promising approach for efficient and accurate systems. Another work (Settles & Craven, 2008) proposes to work with the general active learning strategy of EGL by calculating it as an expectation over the N-best labelings, where the Euclidean norm of each resulting gradient vector is computed. Since discriminative models like Conditional Random Fields are trained using gradient-based optimization, this involves querying the instance which, if labeled and added to the training set, would create the greatest change in the gradient of the objective function. However, it's mentioned that this technique is computationally expensive as it must first perform inference over the possible labelings and then calculate gradients for each candidate label sequence. EGL technique has been used and compared in many works. Multiple Instance Active Learning also considers EGL and compares it with other acquisition techniques. An active learning approach proposed by (Y. Zhang, Lease, & Wallace, 2017) focuses on sentence classification, specifically selecting instances that contain words expected to have the most impact on the embeddings. They calculate the expected gradient length (EGL) concerning the embeddings for each word in the remaining unlabeled sentences. This helps in rapidly learning discriminative and task-specific embeddings. The approach demonstrates effective classification of sentiment in sentences by separating embeddings of words like 'bad' and 'good' quickly. As per this work, one key difference between the EGL approach and the other active learning methods is its focus on word-level embeddings. Instead of considering the entire text or document as a unit for selection, the EGL approach specifically targets individual words within sentences. By selecting instances that contain words expected to have the most influence on the embeddings, the approach aims to achieve discriminative and task-specific embeddings more rapidly. An alternative strategy that leverages the concept of Expected Model Change entails a new active learning strategy for regression tasks called Expected Model Change Maximization (EMCM) (Cai, Zhang, & Zhou, 2013). The motivation behind active learning in regression is to maximize the performance of a learning model while minimizing the cost of data annotation by using as few labeled training examples as possible. The change is estimated using the gradient of the loss function with respect to a candidate example for active learning, inspired by the Stochastic Gradient Descent (SGD) update rule. The paper derives novel active learning algorithms for both linear regression and nonlinear regression within the EMCM strategy. These algorithms are designed to select the most informative examples that lead to significant changes in the model parameters. Another effective method for active learning discussed by (Mussmann et al., 2022) is Expected Error Reduction (EER). It aims to select candidate samples that, on average, minimize the error on an unlabeled set. However, it has not been widely used for modern deep neural networks due to its computational cost as the model needs to be retrained for every candidate sample. So, (Mussmann et al., 2022) tried to reformulate the concept of Expected Error Reduction from a Bayesian active learning perspective deriving a computationally efficient version that can utilize any Bayesian parameter sampling method.

2.2.7 Alternative Approaches in Active Learning

While the aforementioned active learning methods have proven effective, researchers have been exploring alternative approaches further to enhance the efficiency and effectiveness of active learning strategies. One such example could be the Automated Design of Deep AL which refers to the methods that utilize automation to design an Active Learning Query strategy affecting the overall Deep AL performance. Some other active learning-based methods like Fisher Information (Settles & Craven, 2008) focus on active learning with sequence models. Fisher information is a measure of the overall uncertainty about the estimated model parameters. It is a theoretical strategy proposed by (T. Zhang & Oles, 2000) represented as a K × K covariance matrix for a model with K parameters. Another work that discusses a departure from the traditional active learning setting, where only labeled examples are used for model training while unlabeled data is solely used for acquisition was presented in (Siméoni, Budnik, Avrithis, & Gravier, 2020). This work proposes a different approach by incorporating both labeled and unlabeled data throughout the active learning process. It involves utilizing unsupervised feature learning at the beginning and semi-supervised learning at each active learning cycle, incorporating all available data. The use of unsupervised feature learning in active learning is a novel aspect that has not been extensively studied before. Similarly, the exploration of semi-supervised learning in the context of deep learning is relatively scarce, and recent findings regarding its benefits are inconclusive (Rebuffi, Ehrhardt, Han, Vedaldi, & Zisserman, 2020). The main idea is to leverage the additional unlabeled data during model training, similar to how ensemble methods utilize multiple models. By systematically evaluating various acquisition strategies and datasets, the study reveals that incorporating unlabeled data during model training leads to surprising improvements in classification accuracy, surpassing the differences observed between different acquisition strategies. As a result, the research investigates the effectiveness of active learning with smaller label budgets, even with only one labeled example per class. Another similar work that introduced an efficient way to make use of both labeled and unlabelled data with a limited number of labeled samples in multi-class classification tasks was presented in (Rebuffi et al., 2020). It introduces two key ideas where the first idea involves leveraging transfer learning and self-supervision techniques to initialize a robust representation of the data without relying on any labels. By leveraging the knowledge learned from related tasks and self-supervised learning, the model can build a strong initial representation of the data. The second idea presented is a novel self-supervised learning algorithm designed to effectively utilize the pre-trained representation. Recent advancements in deep generative models can also be leveraged to achieve superior results compared to existing approaches in active learning. One such method uses Adversarial Representation Active Learning (Mottaghi & Yeung, 2019) where the approach was to utilize not only labeled data but also unlabeled and generated data to co-train the entire model. By incorporating unlabeled and generated data into the training process, the proposed method achieves higher classification accuracy while using as few labeled samples as possible. It showcases a successful utilization of deep generative models to enhance active learning performance. Another approach called active learning by query synthesis using Generative Adversarial Networks was proposed in (J. Zhu & Bento, 2017). This approach dynamically generates training instances to expedite the learning process and the generated queries are based on the uncertainty principle, although the approach is supposed to adapt to other active learning principles. Later, a model called Semi-supervIsed GeNerative Active Learning (SIGNAL) was designed by (Jiang et al., 2020) to address several challenges like class imbalance, efficiency, and the problem of text camouflage in the task of Chinese text spam detection. It introduces a self-diversity criterion that measures the worthiness of a candidate instance for annotation. This criterion helps identify the most informative samples to be labeled, considering their potential contribution to the learning process.

2.3 Navigating Active Learning Challenges: Strategies for Success

Active learning encounters numerous challenges that necessitate effective solutions for successful implementation. In this section, we will go through a few key challenges faced in real-world scenarios. One of the primary challenges in active learning is query selection (Settles, 2011), which refers to the process of selecting the most informative instances from the unlabeled data pool to query for labeling. The main objective is to choose instances that can provide the most valuable information to the learning algorithm, thereby improving the model's performance. The selection strategy for queries varies depending on factors such as the nature of the problem, available resources, and data characteristics. Often, a combination of different strategies or their variations is employed to achieve a more effective and diverse selection of instances for labeling as per the strategies mentioned above. The ultimate goal is to strike a balance by maximizing the learning impact while minimizing the annotation effort and cost. It requires careful consideration and experimentation to identify the most suitable query selection approach for a specific active learning scenario. Another significant challenge is the cost associated with data annotation. Annotating data can be a labor-intensive and time-consuming process, especially in complex tasks that require domain expertise. Selecting the most informative instances for annotation becomes crucial to maximize the utilization of limited annotation resources. Another challenge is the quality of the labeling process. Human annotators may introduce errors, inconsistencies, or biases while labeling the data. Ensuring the accuracy and reliability of labeled data is essential for training robust and reliable models. Quality control measures, iterative annotation with assistive labels, and reviewer feedback mechanisms can help mitigate labeling errors and maintain high-quality annotated data. Adaptability to changing data distributions is another challenge in active learning. Data distributions may evolve or drift over time, making the initially selected informative instances less representative or informative. Continuous monitoring, retraining, and adaptation strategies are necessary to ensure that the active learning model remains effective and performs well as new unlabeled data becomes available. Additionally, the cold start problem is one of the most faced condition challenges in active learning, particularly when starting with small or no labeled data (Yuan et al., 2020). The model lacks initial knowledge about the data distribution and informative instances, making it difficult to select instances for annotation. Strategies such as domain knowledge incorporation, leveraging pre-existing labeled data from related tasks, or few-shot learning can help overcome the cold start problem.

To address some of these challenges in active learning, researchers have explored the use of deep pre-trained models. These models, when fine-tuned on smaller, task-specific datasets, enable transfer learning and improve model performance while reducing the need for extensive labeled data. In a recent empirical investigation conducted by (Shelmanov et al., 2021), the effectiveness of deep pre-trained models combined with Monte-Carlo Dropout and various Bayesian uncertainty estimation methods was evaluated for Sequence Tagging tasks. The study aimed to assess the potential of fine-tuning pre-trained models and the impact of different Bayesian techniques on model performance. The findings showcased that incorporating these techniques can enhance computational efficiency and facilitate the practical implementation of deep active learning. Another empirical study by (Dor et al., 2020) focused on active learning using BERT-based classification on a large scale. This research employed active learning techniques to boost the performance of BERT models in diverse domains and settings while exploring advanced AL strategies. The objective was to demonstrate the effectiveness of active learning in improving the performance of BERT-based models and their applicability across different domains. These studies emphasize the potential benefits of leveraging deep pre-trained models and active learning techniques in text annotation tasks. By enhancing model performance, simplifying implementation, and reducing the reliance on labeled data, these approaches offer promising avenues for advancing text annotation tools. Overall, the exploration of deep pre-trained models and active learning techniques in text annotation tasks presents an exciting opportunity to address the challenges faced in AL. The findings from these studies shed light on the effectiveness of these approaches and their potential to improve model performance, increase efficiency, and reduce annotation efforts in various text annotation scenarios.

2.4 Summary

This chapter talks about various active learning strategies, highlighting the challenges and presenting effective solutions. Text annotation is a crucial step in many natural language processing applications, but manual labeling can be time-consuming and expensive. Active learning techniques can effectively address these challenges by intelligently selecting the most informative instances for annotation, reducing the annotation effort while maintaining high-quality models. This chapter also explores the integration of active learning with other techniques to further enhance performance like combining active learning with semi-supervised learning, where unlabeled data is leveraged in conjunction with labeled data for training. Additionally, the use of transfer learning, pretraining models, and ensemble methods are discussed as strategies to improve active learning performance. Furthermore, it highlights recent advancements in deep active learning, which utilize deep neural networks to address different challenges. It talks about several techniques categorized into various strategies and some techniques may fall into multiple categories depending on the nature of the technique. Overall, the section emphasizes the importance of active learning strategies and their potential to mitigate annotation challenges. By intelligently selecting informative instances for annotation, active learning enables the creation of high-quality models with reduced annotation efforts. The exploration of different active learning strategies and their integration with other techniques offers valuable insights into the development of efficient data annotation strategies.

Chapter 3

Synergizing Active Learning and Few-Shot Learning

3.1 Overview

In today's world, critical infrastructure systems such as security, health, and transportation rely heavily on a large number of terminal devices that generate a significant amount of data. Storing and processing this data remains a significant challenge due to difficulties in extracting useful labeled data, and lack of model generalization when applying the same models across diverse domains resulting in sub-optimal performance in new and different settings. Existing annotation tools and systems still have a huge room for improvement, particularly in scenarios where new classes need to be added, or the model needs to adapt to changing data distributions. One promising direction that is gaining traction to address these challenges is Few-shot Learning (FSL). It aims to address challenges related to data distribution, future task generality, and feature reuse sensitivity as well. FSL involves building an accurate model using a minimal set of samples. The primary objective of FSL is to learn how to learn, which is accomplished by finding similarities and differences between samples using a similarity function. We will explore some key concepts in the field of FSL that play a crucial role in advancing the research (Song, Wang, Mondal, & Sahoo, 2022). These few-shot concepts are integral for enabling models to effectively learn from a limited number of labeled examples. The field of FSL has witnessed significant advancements in recent years, offering innovative approaches to tackle the challenge of learning from limited labeled examples. This chapter aims to provide a comprehensive exploration of the various concepts that underpin the foundations of FSL. By delving into these concepts, we can gain a deeper understanding of the strategies and techniques that empower models to generalize effectively and make accurate predictions, even in scenarios with scarce labeled data.

The main learning objective of FSL is to classify unseen classes with limited labeled data. FSL can be defined in multiple ways, but the most widely accepted formal definition was presented in (Y. Wang, Yao, Kwok, & Ni, 2021). According to this definition, FSL is a type of machine learning problem, which is understood as a computer program that can learn from experience E in order to improve its performance on certain classes of tasks T, as measured by performance measure P. It is important to note that in FSL, the amount of experience E available is limited. Numerous surveys and studies have aimed to investigate various techniques in FSL and develop a taxonomy for these techniques. Some of these efforts have attempted to differentiate the techniques based on different learning behaviors (Shu, Xu, & Meng, 2018), while others rely on different categories of models, like Generative and Discriminative (Lu, Gong, Ye, & Zhang, 2020). Some have taken a broader perspective based on prior knowledge with respect to data, model, and algorithm (Y. Wang et al., 2021). More recently, another survey by (Song et al., 2022) proposed a taxonomy based on the challenges faced by the community in few-shot learning. Based on the existing taxonomies, one possible way to categorize FSL techniques is to first roughly divide them based on their evolution period, then based on the models types, and finally based on the challenges faced by users. Research in FSL has undergone significant evolution over time, with two distinct periods marked by a significant breakthrough in 2015 by (Koch, Zemel, Salakhutdinov, et al., 2015). This breakthrough introduced the first combination of deep learning techniques with FSL problems. During the evolution of FSL, approaches were predominantly classified as either generative or discriminative model-based. In the latter period of development, discriminative-based models dominated generative-based models. As a result, FSL research has become increasingly sophisticated, enabling the development of more effective models for solving complex real-world problems.

Prior to the milestone in 2015, all FSL solutions were based on non-deep learning methodologies or techniques. The congealing algorithm (Miller, Matsakis, & Viola, 2000) is one of the earliest pioneers in the study of learning from very few samples. Then the term ”one-shot learning” in the context of few-shot appeared in the paper (Fei-Fei, Fergus, & Perona, 2003) which proposed a Bayesian approach to unsupervised one-shot learning of object categories. During the earlier period of FSL research, the majority of popular approaches were based on generative models. These methods aimed to estimate either the joint distribution P(X, Y) or the conditional distribution P(X/ Y) given very few training samples with supervision, and then make predictions for test samples using Bayesian decision theory. While there were also several discriminative model-based FSL approaches during this period such as (Fink, 2004), (Wolf & Martin, 2005a), and (Bart & Ullman, 2005), they were not as popular. Discriminative model-based FSL approaches aim to learn a conditional distribution P(Y/ X) directly from a single observed sample, which allows for direct probability prediction. The generative models learn the underlying probability distribution of the data, while discriminative models directly learn the decision boundary that separates the classes. Generative models can be more flexible in modeling complex distributions, but discriminative models are generally more efficient and can perform better in high-dimensional spaces with limited data.

Despite all these efforts, the FSL research during this period progressed slowly. The emergence of deep learning (Krizhevsky, Sutskever, & Hinton, 2012) (Wolf & Martin, 2005b), led many FSL researchers to shift their focus to deep learning models. The year 2015 witnessed a groundbreaking development in the field of FSL, when (Koch et al., 2015) pioneered the integration of deep learning with FSL through the introduction of Siamese CNNs. This innovative approach enabled the network to learn a class-agnostic similarity metric for pairwise samples, thus revolutionizing the way FSL is approached and studied. A few generative techniques like Sequential Generative Models (Rezende, Mohamed, Danihelka, Gregor, & Wierstra, 2016) and Neural Statisticians (Edwards & Storkey, 2016) were also introduced in this era but the majority of FSL techniques were discriminative-based. Following the introduction of deep neural networks in FSL research, subsequent approaches have fully leveraged their advantages in end-to-end model optimization and feature representation to tackle FSL problems from various perspectives. One of the most effective ways to enhance active learning and optimize human effort is by integrating it with few-shot fine-tuning via transfer learning. By integrating few-shot fine-tuning via transfer learning with active learning, we can significantly enhance the overall effectiveness and efficiency of the active learning process. This integration offers several key benefits, such as improving the optimization of the learning model and reducing the amount of human effort required in the active learning pipeline. With few-shot fine-tuning, the model can leverage prior knowledge and generalize from a small amount of labeled data, enabling it to quickly adapt and make accurate predictions on new, unseen samples. This reduces the dependency on large labeled datasets and extensive human annotation, thereby decreasing the overall human effort needed for active learning.

3.2 Key concepts in Few-Shot Learning

This section will shed light on the key concepts of FSL like the N-way-K-shot problem, meta-learning, metric learning, and transfer learning, and their significance in advancing the field of few-shot learning. Through this exploration, we aim to uncover the principles and methodologies that drive the success of few-shot learning algorithms and inspire further advancements in this exciting research area. Two key concepts related to FSL are the N-way-K-shot problem and the cross-domain FSL (Song et al., 2022). The N-way-K-shot problem is a common formulation used in FSL. It refers to the scenario where a model is trained on a small support set, which consists of a limited amount of labeled data. This support set is used to provide reference information for the model's testing phase. The query set, on the other hand, represents the actual tasks on which the model needs to make predictions. Importantly, the classes in the query set do not appear in the support set. The notation N-way-K-shot signifies that the support set contains N categories, each with K samples, resulting in a total of N * K samples for the task. For instance, N-way-1-shot represents one-shot learning, where there is only one labeled example per category, and N-way-0-shot corresponds to zero-shot learning, where there are no labeled examples for the target categories in the support set. On the other hand, the concept of cross-domain originates from transfer learning, which involves transferring knowledge from a source domain to a target domain. Cross-domain FSL combines the characteristics of both cross-domain learning and FSL. It deals with the challenges of transferring knowledge and learning from one domain to another, where domain gaps and differences may exist.

This integration of cross-domain and FSL presents a challenging research direction that has gained attention recently. To further understand how we can deal with various challenges in FSL (Song et al., 2022), let us now dive deeper into the FSL concepts like Meta-Learning, Metric Learning, and Transfer Learning.

Meta-Learning refers to the ability of a model to quickly adapt to new tasks with limited data by building mappings from known tasks to target models in previously unseen tasks (Hochreiter, Younger, & Conwell, 2001). In meta-learning, the algorithms are designed to generate a ”base” model that can efficiently adapt to new tasks through a process like fine-tuning. This base model is created by training it on a small set of sample tasks drawn from a larger distribution of tasks. By using this training approach, the model is able to quickly generalize to new tasks it has not seen before, making it a powerful tool for applications that require rapid learning and adaptation (Hochreiter et al., 2001) (Goldblum, Fowl, & Goldstein, 2020). The task and data are both sampled, and the meta-knowledge is summarized from different tasks to enable fast integration of unseen tasks at a lower cost. In recent years, there has been a surge of interest in meta-learning-based FSL techniques, with several innovative approaches emerging. These include Matching Nets (Vinyals, Blundell, Lillicrap, Kavukcuoglu, & Wierstra, 2016), MAML (Finn, Abbeel, & Levine, 2017), Meta-Learner LSTM (Ravi & Larochelle, 2017), Prototypical Nets (Snell, Swersky, & Zemel, 2017), and LGM-Nets (Li et al., 2019), among others. Meta-learning strategies have become the dominant approach for FSL, and these advanced techniques have been directly applied or improved to solve a wide range of problems in computer vision, natural language processing, audio and speech, data analysis, robotics, and other fields. It is not a distinct FSL model; instead, it is a cross-task learning strategy at a high level. The progress made in meta-learning and FSL in recent years has opened up exciting new avenues for research and practical applications.

Metric Learning is a branch of the meta-learning-based FSL approach and it is sometimes referred to as a Learn to Measure technique (Musgrave, Belongie, & Lim, 2020) (Lu et al., 2020). It is an embedding-based technique that focuses on learning a metric space that can effectively capture the similarity or distance between samples, thereby enabling accurate classification of previously unseen samples. This involves learning a metric that can effectively distinguish between different classes using only a limited amount of labeled data. Metric learning has extensively been studied in the literature, with several notable works being developed over the years. Among these, Neighborhood Components Analysis (NCA) (Goldberger, Roweis, Hinton, & Salakhutdinov, 2004) is a popular approach that learns a Mahalanobis distance to maximize the leave-one-out accuracy of K-nearest neighbors (KNN) in the transformed space. Another notable work is an extended version of NCA by (Salakhutdinov & Hinton, 2007), which uses a neural network to perform the transformation. Another popular approach is Prototypical Networks (Snell et al., 2017) which are very similar to the non-linear extension of NCA (Salakhutdinov & Hinton, 2007), but they form a softmax directly over classes, unlike non-linear NCA. Their approach is also similar to the nearest class mean approach (Mensink, Verbeek, Perronnin, & Csurka, 2013) where it utilizes neural networks to non-linearly embed points and handles few-shot scenarios. The authors also highlight that Prototypical Networks learn a non-linear embedding in an end-to-end manner, producing a non-linear classifier that still only requires one prototype per class, and the approach generalizes well to other distance functions, specifically Bregman divergences. Siamese Neural Network is a metric-based learning model that employs a discriminator with a set of positive or negative sample pairs as input to the model. During the inference stage, the model evaluates the similarity of the incoming data. The concept of Siamese networks was first introduced in (Bromley, Guyon, LeCun, Säckinger, & Shah, 1993), where they were used to identify signature forgeries using dual identical sub-networks. The sub-networks were trained by extracting features from two given signatures and measuring the distance between the two using the joining neuron. This approach enabled the creation of highly discriminative features that could distinguish between genuine and forged signatures. The extracted feature vectors were then compared to the existing stored feature vectors to verify the signatures. Since their inception, Siamese networks have been successfully applied in various applications such as face recognition, image retrieval, and more. The effectiveness of Siamese networks lies in their ability to learn a similarity metric that can generalize well across different data distributions and enable accurate classification with limited labeled examples. The input text and its labels are encoded independently so the label vectors are pre-computed and the similarity of the two vectors can be measured using various similarity functions like cosine similarity. One such work which uses cosine similarity to derive semantically meaningful sentence embeddings using a modification of BERT model (Devlin, Chang, Lee, & Toutanova, 2019) with siamese network and triplet network structures was presented by (Reimers & Gurevych, 2019). These methods are just a few examples of the vast body of work in metric learning and have been shown to be effective in various tasks such as image classification and object detection.

Transfer Learning is a machine learning technique where knowledge gained from training a model on one task is leveraged to improve the performance of a related task or domain with minimal additional training data. The goal is to transfer the learned features, representations, or knowledge from one task or domain to another, where the target task or domain has a different but related distribution of data (Zhuang et al., 2021). It is a technique that uses pre-training and fine-tuning to extract prior knowledge from large-scale auxiliary datasets (Bozinovski & Fulgosi, 1976). This approach can be particularly useful in the context of FSL, as it allows models to leverage prior knowledge from large datasets to improve performance on specific tasks. Through a continuous accumulation of a priori knowledge from large-scale auxiliary datasets, transfer learning enables models to easily transfer knowledge from the source domain to a similar target domain. The pre-training stage involves extracting high-dimensional feature vectors through a feature extractor, while the fine-tuning stage is focused on making minor adjustments to the initial parameters of the pre-training to optimize performance on the target task. Recent advances in language model fine-tuning, such as Text-to-Text approaches, have achieved state-of-the-art few-shot performance through techniques like in-context learning (ICL). These advancements have improved the ability of language models to adapt to new tasks with limited training data[1]. With the scaling of model size and corpus size, large language models (LLMs) have shown remarkable progress in ICL (Dong et al., 2023). In numerous studies, LLMs have demonstrated their capability to perform a variety of complex tasks through ICL, including mathematical reasoning problems (Wei et al., 2022). It relies on the principle of analogical reasoning, where an LLM can infer new knowledge by relating it to previous knowledge and reasoning based on the analogy between them. This approach allows LLMs to acquire new knowledge quickly and efficiently, making them a promising tool for various applications. However, this technique has several challenges associated with it like its inability to distill to smaller models, issues in improving and updating knowledge in LLMs, and various other problems related to using

[1]https://towardsdatascience.com/sentence-transformer-fine-tuning-setfit-outperforms-gpt-3-on-few-shot-text-classification-while-d9a3788f0b4e

ICL for data annotation, augmentation, pruning, and adversarial data generation (Dong et al., 2023). Another very popular technique called T-Few (H. Liu et al., n.d.) is a parameter-efficient few-shot learning recipe that achieves strong performance on novel tasks while updating only a tiny fraction of the model's parameters. It is based on the T0 model (Sanh et al., 2022) and `(IA)^3` method which is a Parameter Efficient Fine-Tuning (PEFT) method introduced in this paper that re-scales inner activations with learned vectors. T-Few also uses two loss terms in addition to a standard cross-entropy loss, which helps the model to lower the probabilities for an incorrect choice and account for the length of different answer choices using an unlikelihood loss and a standard softmax cross-entropy loss, respectively. The technique was evaluated using the RAFT benchmark (Alex et al., 2021) and achieved super-human performance while outperforming prior submissions by a large margin. The authors also performed computational cost analysis which shows that T-Few uses over 1,000× fewer FLOPs during inference than few-shot ICL with GPT-3. Overall, T-Few provides a new perspective on how to effectively perform few-shot learning with LLMs on classification tasks.

In addition to techniques like PEFT (H. Liu et al., n.d.), other FSL methods like Pattern Exploiting Training (PET) called ADAPET(Tam, Menon, Bansal, Srivastava, & Raffel, 2021), or prompt-free techniques like PERFECT (Mahabadi et al., 2022) have shown impressive results in label-scarce scenarios. However, these methods are challenging to employ due to their high variability from manually crafted prompts and typically require billion-parameter language models to achieve high accuracy (Tunstall et al., 2022). To address these challenges, a Sentence Transformer Fine-tuning (SETFIT) method was proposed by (Tunstall et al., 2022). SETFIT is an efficient and prompt-free framework for few-shot fine-tuning of Sentence Transformers. It fine-tunes a pre-trained sentence transformer in a contrastive Siamese manner on a small number of text pairs to generate rich text embeddings, which are used to train a classification head. The approach is prompt-free and requires no verbalizers, making it much simpler to implement and less prone to variability. During inference, the fine-tuned transformer encodes an unseen input sentence and produces a sentence embedding, based on this sentence embedding the classification head then produces the class prediction. It achieves high accuracy while requiring fewer parameters than existing techniques. SETFIT underwent evaluation on various benchmarks, including RAFT, where it outperformed the human baseline in 7 out of 11 tasks. It is over 30 times smaller than the T-FEW model and does not require manual prompt crafting. Additionally, experiments conducted on pre-defined benchmarks showed that SETFIT is considerably faster during both training and inference compared to comparable approaches such as T-FEW, ADAPET, and PERFECT. Furthermore, SETFIT requires much smaller base models to achieve high performance without external computing, making it an efficient and effective few-shot learning method that can be utilized well for an efficient text annotation platform. In recent years, FSL-based transfer learning techniques have been applied not only to text classification tasks but also to sequence labeling tasks such as named entity recognition. To support research in this area, (N. Ding et al., 2021) created a large-scale human-annotated named entity recognition dataset for FSL. Additionally, they performed experiments using existing state-of-the-art models like BERT-Tagger (Devlin et al., 2019), ProtoBERT (Snell et al., 2017), NNShot, and StructShot (Yang & Katiyar, 2020b) on this dataset to evaluate their performance. Another work proposed by (Müller, Pérez-Torró, & Franco-Salvador, 2022) focuses on the problem of building text classifiers when there is limited or no training data available, which is commonly referred to as zero and few-shot text classification. Traditionally, neural textual entailment models have been used for this purpose and have shown strong results across various tasks. However, the researchers propose an alternative approach using Siamese Networks, which embed texts and labels, and demonstrate that with proper pre-training, these models can provide competitive performance. One significant advantage of using Siamese Networks is the reduced inference cost, as it is constant in the number of labels rather than linear. They also introduced label tuning, a computationally efficient approach that enables adapting the models in a few-shot setup by only modifying the label embeddings. Although label tuning may yield lower performance compared to model fine-tuning, it offers the architectural advantage of sharing a single encoder across multiple tasks.

Sometimes it can be challenging to apply transfer learning when a large domain shift occurs between the source and target domains, as the learned knowledge may not be directly applicable to the new domain. So, we can use concepts like Cross-domain few-shot learning (Tseng, Lee, Huang, & Yang, 2020) which is a subfield of few-shot learning that deals with the problem of generalizing knowledge from a few samples of new classes in a target domain, when those classes do not exist in the source domain. In this scenario, the model has to learn to recognize and classify new classes with only a few examples in a domain that is different from the one it was trained on. It combines the challenges of both transfer learning and few-shot learning. There is no intersection of classes between the source and target domains, and the available sample size for each class in the target domain is extremely small. This makes it difficult for the model to generalize well to new classes and domains. To further enhance the discussion on cross-domain few-shot learning, it's important to note that it involves not only distinguishing domain-irrelevant features but also employing domain adaptive techniques using transfer learning. This includes identifying and transforming features (Tseng et al., 2020) that are most relevant to the target domain, as well as creating auxiliary datasets to help the model learn new classes with limited samples. By leveraging transfer learning and domain adaptation, the model's generalization capability can be improved, even in the face of large domain shifts between the source and target domains.

3.3 Fusion of Active Learning and Few-Shot Learning

3.3.1 Synergistic potential of combining AL and FSL

The combination of AL and FSL presents a promising opportunity for synergistic advancements in model development. By integrating these two approaches, researchers seek to enhance the efficiency and effectiveness of the annotation process and improve overall model performance. AL plays a crucial role in the annotation process by intelligently selecting the most relevant instances for labeling. This reduces the reliance on extensive manual annotation and saves valuable time and resources, especially in situations where labeling large datasets is costly or time-consuming. The integration of FSL techniques with AL further strengthens the annotation process. FSL models, designed to adapt quickly to new tasks with limited labeled data, can contribute valuable insights and knowledge transfer to AL. Leveraging FSL models, the AL process can identify critical data points that maximize model performance with minimal labeling efforts. Additionally, the combination of AL and FSL offers the potential for improved model generalization and performance. FSL models excel at learning from a few labeled examples and generalizing well to new, unseen instances. AL complements this by actively selecting diverse and informative samples, refining the model's adaptability and generalization across various tasks and domains. Chapter 2 discusses how FSL mitigates challenges in AL, such as model adaptability to changing data distributions and cold start problems.

Existing systems like (Margatina, Vernikos, Barrault, & Aletras, 2021) start with a small set of labeled examples, ranging from 100 to several thousand, and systems like (Grießhaber, Maucher, & Vu, 2020) work with few-shot scenarios, typically involving less than a thousand training examples, but they still rely on a seed set of labeled data. However, in real-life scenarios, it is often impractical to assume the availability of labeled data from the outset. This issue becomes particularly relevant in settings where obtaining labeled data is costly, time-consuming, or impractical. Therefore, addressing the Cold Start challenge necessitates the development of novel approaches that effectively learn from limited or even zero-labeled data. In this context, few-shot and zero-shot learning paradigms play a crucial role. Research works such as (Müller, Pérez-Torró, Basile, & Franco-Salvador, 2022) and (Yuan et al., 2020) focus on exploring the domains of active learning and few-shot learning for the relatively unexplored cold-start scenario. These works aim to develop innovative strategies that enable efficient learning from limited or zero-labeled data, further advancing the integration of AL and FSL.

3.3.2 AL-FSL Integration in NLP

The integration of Active Learning and Few-Shot Learning techniques in the field of Natural Language Processing has the potential to unleash a new level of power and efficiency in various NLP models. By combining the strengths of Active Learning, which focuses on selecting the most informative instances for labeling, and FSL, which enables effective generalization with limited labeled data, researchers can harness the benefits of both approaches simultaneously. This integration allows for the intelligent selection of labeled instances during active learning iterations, while also leveraging the knowledge and patterns captured through FSL to enhance the model's ability to generalize to unseen classes or tasks. This synergy between AL and FSL opens up exciting possibilities for advancing the capabilities of NLP models, enabling them to achieve higher performance and efficiency in various NLP tasks. One of the biggest challenges of training text classification models in natural language processing is the need for a large number of labeled examples. Therefore, (Müller, Pérez-Torró, Basile, & Franco-Salvador, 2022) suggested using the concepts of FSL and AL as two research directions to address this problem. The work presented combines these two approaches into a platform called FASL, which enables the iterative and efficient training of text classification models. The researchers investigate different active learning methods within the FSL setup to determine the most effective approach. Furthermore, they develop a model that predicts when to stop annotating data, which is particularly important in FSL scenarios where a large validation set is not available. It focuses on investigating the efficacy of AL techniques on both balanced and unbalanced datasets, shedding light on the challenges and potential solutions in this unique setting. Another work that discusses the challenges and limitations of FSL in the context of text classification was presented by (Köksal, Schick, & Schütze, 2022). It highlights the high variance observed across different sets of few-shot examples (data selection) and across different fine-tuning runs, which hinders fair comparisons and makes FSL unreliable for real-world applications. To address these issues, (Köksal et al., 2022) proposed two concepts. First, they introduce novel ensembling methods that significantly reduce the variability across different fine-tuning runs, leading to more stable few-shot learning. Second, they present a new active learning criterion for data selection, specifically tailored towards prompt-based learning. Then these methods are combined into a framework called MEAL (Multiprompt finetuning and prediction Ensembling with Active Learning). Similarly, a library by (Schröder, Müller, Niekler, & Potthast, 2023) was designed to be easy to use and provide pre-implemented state-of-the-art query strategies, some of which leverage GPU capabilities. It offers standardized interfaces that allow users to combine different classifiers, query strategies, and stopping criteria, enabling a flexible and efficient development of active learning experiments and applications. The small-text library integrates well-known machine learning libraries such as scikit-learn, PyTorch, and Hugging Face transformers, making various classifiers and query strategies accessible. The library also supports optional installations for GPU usage. There are also some tools like Argilla[2] which empower users to build robust language models by facilitating faster data curation through a combination of human and machine feedback. Unlike previous studies that primarily focused on low inter-task variance in image domains, (Yu et al., 2018) addresses the more realistic scenario of diverse tasks in natural language processing. The challenge lies in the fact that a single metric is inadequate to capture the complex task variations in this domain, making existing metric-based algorithms ineffective. To address this limitation, (Yu et al., 2018) propose an adaptive

[2] https://docs.argilla.io/en/latest/

metric learning approach that automatically determines the optimally weighted combination of metrics obtained from meta-training tasks for new few-shot tasks. Extensive quantitative evaluations conducted on real-world sentiment analysis and dialog intent classification datasets demonstrate that the proposed method outperforms state-of-the-art few-shot learning algorithms in terms of predictive accuracy. Another work by (Z. L. Zhu, Yadav, Afzal, & Tsatsaronis, 2022) discusses the challenges in domain-specific applications that require annotating and labeling large volumes of unstructured textual data. To address this problem, the paper proposes a novel approach that combines active learning and meta-learning. The active learner is initialized with meta-learned parameters obtained through meta-training on tasks similar to the target task. The approach utilizes the pre-trained BERT as the text-encoder and meta-learns its parameters using LEOPARD, an extension of the model-agnostic meta-learning method. LEOPARD generates task-dependent softmax weights to enable learning across tasks with varying numbers of classes. The effectiveness of the proposed method is demonstrated through experiments on five natural language understanding tasks and six datasets, using five different acquisition functions. The results show that the approach outperforms the baseline, especially when closely related tasks were present during meta-learning. Additionally, the study reveals that better performance with fewer labeled samples leads to better performance when larger acquisition batches are used. The ablation study indicates that active learning with only the meta-learned weights is beneficial while adding the meta-learned learning rates and generating the softmax has negative consequences for performance. One important approach introduced in (Asghar, Poupart, Jiang, & Li, 2017) proposes an online, end-to-end, neural generative conversational model for open-domain dialogue. It is trained using a unique combination of offline two-phase supervised learning and online human-in-the-loop active learning. The model promotes the generation of semantically coherent, relevant, and interesting responses and can be trained to adopt diverse moods, personas, and conversation styles. The hamming-diverse beam search mechanism is used for response generation and one-character user feedback is provided at each step. The model can be used to create agents with customized backgrounds and characters. The experiments show that the model outperforms existing models in generating semantically relevant and interesting responses.

3.3.3 AL-FSL Integration in Other Task Domains

Harnessing the potential of AL and FSL integration in various task domains opens up exciting possibilities for advancing machine learning capabilities. By combining the benefits of AL and FSL, we can overcome the limitations of traditional learning approaches and achieve remarkable results in tasks requiring limited labeled data. In this section, we explore the transformative impact of AL-FSL integration in different task domains, shedding light on its potential applications and the key advantages it brings to the table. A concept of using a continual learning agent was introduced in (Ayub & Fendley, 2022) where the problem of Few-Shot Continual Active Learning (FoCAL) was addressed. In FoCAL, a Continual Learning (CL) agent is given unlabeled data for a new or previously learned task in each increment, but it only has a limited budget for labeling. The paper builds upon the literature on continual learning and active learning to develop a framework that enables the CL agent to learn new object classes with only a few labeled training examples. The framework utilizes a uniform Gaussian mixture model (GMM) to represent each object class and employs pseudo-rehearsal to mitigate catastrophic forgetting. Additionally, uncertainty measures on the Gaussian representations of previously learned classes are used to identify the most informative samples for labeling in each increment. The proposed approach is evaluated on the CORe-50 dataset and a real humanoid robot for object classification, demonstrating state-of-the-art results and the ability of the real robot to continually learn new objects in a real environment with limited labeling supervision. Another novel approach to tackle the problem of few-shot learning is by examining it through the lens of inference on a partially observed graphical model which was introduced in (Satorras & Estrach, 2018). They construct a graphical model using a collection of input images, where the labels can be either observed or unknown. The authors combine traditional message-passing inference algorithms with neural-network counterparts to create a graph neural network architecture. This architecture offers a generalized framework that encompasses several existing few-shot learning models and the proposed framework can be extended to incorporate variants of few-shot learning, such as semi-supervised or active learning. In the active learning setup, the learner has the capability to request labels from a sub-collection of unlabeled samples and the goal is to investigate the improvement in performance compared to the previous semi-supervised setup and achieve similar performance as the one-shot learning setting. The GNN is trained using a Softmax attention mechanism over the unlabeled nodes of the graph. A neural network-based function is applied to each unlabeled vector node to map it to a scalar value, and a Softmax operation is performed over these scalar values to determine the most informative label to query for classification. The experiments in (Satorras & Estrach, 2018) follow the q-shot, K-way setting, where q denotes the number of labeled samples per class and K represents the number of classes. Overall, the approach utilizes active learning to improve the performance of the GNN in the few-shot learning and semi-supervised learning settings by selectively querying informative labels for unlabeled samples. Later, the combination of reinforcement learning and one-shot learning was explored by (Woodward & Finn, 2017) to improve classification tasks. In this work, the authors introduce a classification task where a stream of images is presented, and at each time step, the model must decide whether to predict a label or pay for the correct label. The goal is to train a recurrent neural network-based action-value function that learns when and how to request labels effectively. By designing an appropriate reward function, the model can achieve higher prediction accuracy compared to a similar model trained purely with supervised learning. Furthermore, the model can also trade prediction accuracy for fewer label requests, making it more efficient in terms of resource utilization. Another work that mitigates the limitation of limited availability of labeled data in few-shot learning was presented in (Pezeshkpour, Zhao, & Singh, 2020). They discuss the combination of few-shot learning and active learning. Few-shot learning aims to acquire knowledge about new concepts using only a small number of labeled samples. Active learning, on the other hand, involves the deliberate selection of informative samples to improve model performance. They investigated the effectiveness of actively identifying informative samples in the context of few-shot learning. It finds that while active learning approaches are beneficial for regular classification tasks with larger amounts of labeled data, these benefits do not reliably extend to few-shot learning tasks. The paper introduces two active methods, Single-Instance-Oracle and Batch-Oracle, which assume access to labels of the unlabeled pool and the test set. These methods serve as ”upper bounds” to characterize the best possible performance of active few-shot learning. The findings suggest that actively selecting instances does not offer significant room for improving few-shot models.

3.4 Summary

This chapter discussed various concepts of few-shot learning and explores its combination with active learning to develop an effective annotation tool. Few-shot learning addresses the challenge of training models with limited labeled data, while active learning focuses on iteratively selecting informative instances for model training. By combining these two approaches, researchers aim to enhance the performance of annotation tools. The chapter delves into various techniques and methodologies used in the integration of few-shot learning and active learning. It explores the benefits of leveraging pre-trained models, transfer learning, and self-supervised learning to initialize representations of data without relying heavily on labeled data. The chapter also introduces novel algorithms and criteria for active learning in the context of few-shot learning, enabling the selection of informative samples for annotation. The findings highlight the potential of this integrated approach in developing annotation tools with improved accuracy and efficiency, ultimately facilitating faster and more effective model development.

Chapter 4

Comparative Framework - Text Annotation Tools

4.1 Overview

Despite the availability of several annotation tools, many of them are limited in terms of efficiency and model adaptation. While some modern tools such as Label Studio (Tkachenko, Malyuk, Holmanyuk, & Liubimov, 2020-2022), Label Sleuth (Shnarch et al., 2022), Light Tag (Perry, 2021), Potato (Pei et al., 2022), and Prodigy (*Prodigy*, 2017) offer certain functionalities related to auto-annotation, active learning, and model training, they still do not provide a comprehensive platform that can adapt well to dynamically added labels and may undergo problems like cold-start. Furthermore, these tools are not fully equipped to provide comprehensive insights about their data. Hence, there is a need for a holistic platform that supports assisting auto-labels, a fast adaptation mechanism of data using concepts like few-shot learning, and gradual improvement of annotations over iterations using active learning and fine-tuning.

In this chapter, we present a novel approach to compare various systems working on the annotations based on a comprehensive comparative framework that can be defined using a set of criteria. While existing surveys such as (X. Qiu et al., 2020) and (Ren et al., 2021), compared various annotation systems working in a diverse field of study, they missed a few important technical aspects. Some prior research, including the survey of annotation tools for biomedical literature (Neves &

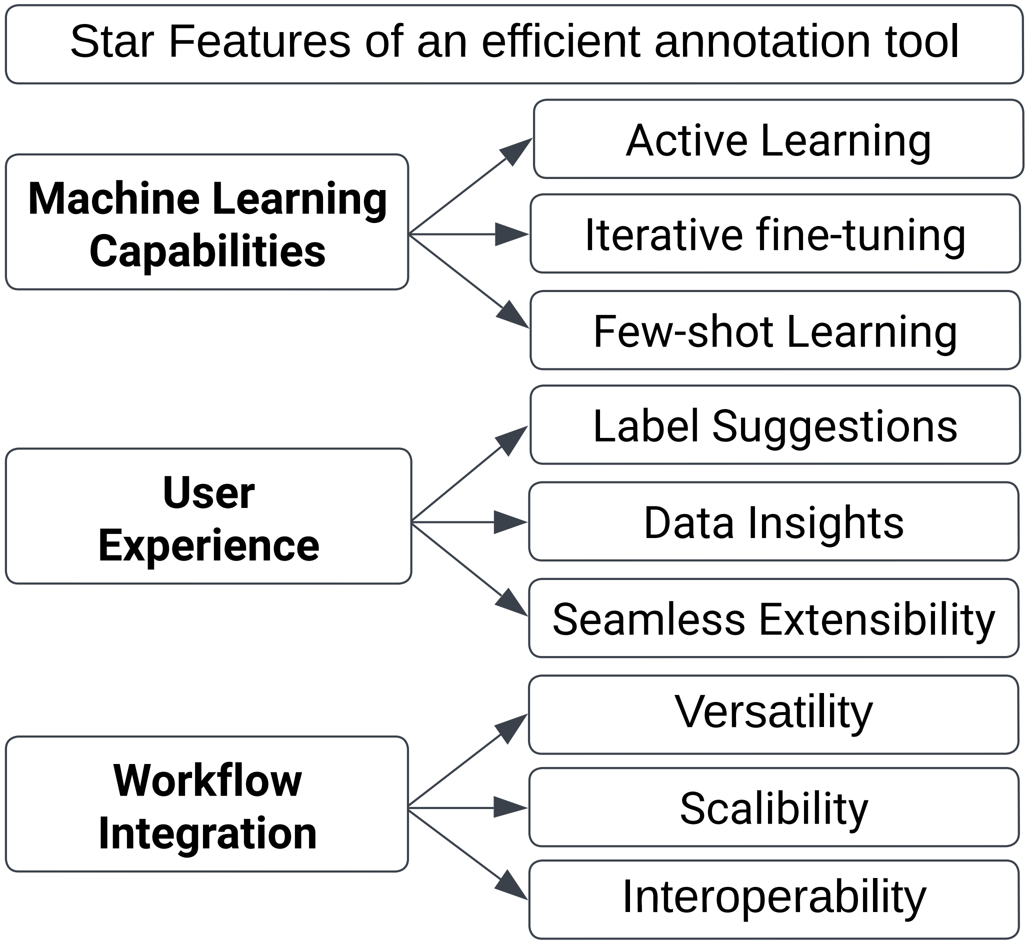

Figure 4.1: A Comparative Framework to differentiate between different auto-annotation-based systems based on various characteristics

Leser, 2014), has evaluated tools based on criteria within categories such as basics, publication, system properties, data, and functionalities. However, these evaluations have lacked a comprehensive consideration of key technical aspects such as efficiency, seamless integration, scalability, and configurability. This study aims to fill this gap by emphasizing crucial technical aspects and providing a thorough comparison of existing annotation systems. To address this issue, we propose to define a set of criteria that considers both the technical aspects and other systematic desirable characteristics of annotation systems. Our primary objective is to evaluate these systems based on a well-defined set of criteria. Throughout this evaluation process, we will carefully consider three critical characteristics, namely: Machine Learning Capabilities, User Experience, and Workflow Integration. Our approach to comparing the annotation systems builds upon the existing surveys and studies in the field but goes further by taking into account the technical aspects that are critical to the success of an annotation system. By defining these three main characteristics, we aim to provide a comprehensive and systematic framework that can be used to evaluate different systems working on the annotation problem.

Existing systems can be differentiated based on the fine-grained features from the above-mentioned three main desirable characteristics as illustrated in Figure 4.1. Let's explore this comparison between several existing systems as showcased in Table 4.1 and explain them in detail.

4.2 Features of the Comparative Framework

4.2.1 Machine Learning Capabilities

This criterion refers to the ability of a system to incorporate machine learning-based capabilities to effectively learn and make predictions on a given set of data. This includes the data used for fine-tuning a model, the model training process, and the ability to adapt to new data. It encompasses the ability of the system to learn from data, improve over time, and make accurate predictions based on that learning. The main goal is to enable the system to accurately annotate data with limited human input. We can consider various features like active learning, few-shot learning and iterative model training to further understand this criterion.

Active Learning: An Active Learning-based Data Acquisition enables an annotation tool to acquire a focused, limited amount of training data in a cost-effective manner. This technique actively selects the most informative examples from the available data pool to be annotated and used for model training, reducing the requirement for large annotated datasets, which can be resource-intensive to acquire. The key advantages of this technique are twofold. First, by focusing on the most valuable examples, the model is trained on relevant data, resulting in improved accuracy and precision of annotations. This targeted approach ensures that the model learns from the most critical data points, enhancing its performance and generalization capabilities. Second, the reduction in the requirement for large amounts of annotated data translates into significant cost and time savings. This approach is a cost-effective solution for large-scale annotation projects with limited resources.

Iterative Model fine-tuning: The significance of model training in annotation tools lies in its ability to derive knowledge from a pre-annotated dataset and utilize it to make predictions on new, previously unseen data. The performance of the training process and the quality of the training data play a pivotal role in determining the accuracy and dependability of the tool's predictions. Improper training of the model or the usage of non-representative training data can result in erroneous and potentially damaging predictions made by the auto-annotation tool. Thus, a thorough and systematic training process is imperative for guaranteeing the efficiency and reliability of the annotation tool. This features focuses on rectifying potential biases, inconsistencies, or labeling errors by providing a way for the human-in-the-loop to give feedback before every model fine-tuning.

Few-shot Learning: As discussed in previous chapters, few-shot learning plays a pivotal role in addressing the challenges of adapting to new data and making accurate predictions over time. Unlike traditional machine learning approaches that heavily rely on extensive labeled datasets, few-shot learning empowers models to generalize effectively from only a limited number of labeled examples. This ability to learn from scarce data is crucial in dynamic environments where new data constantly emerges which is very common in annotation tasks. As the model accumulates knowledge and experience over time, it becomes more adept at recognizing patterns and adapting to novel scenarios. This adaptability is particularly valuable when facing concept drift, where the underlying data distribution changes. By continuously updating its knowledge through few-shot learning, the model can maintain its performance and relevancy, ensuring accurate predictions even as the data landscape evolves. Ultimately, few-shot learning serves as a powerful tool for building robust and intelligent systems capable of learning from sparse data, thereby facilitating continuous improvement and reliable performance over time.

4.2.2 User Experience

In the context of annotation tools, user experience encompasses the overall satisfaction of annotators while utilizing the tool. This includes various aspects, such as the intuitiveness of the user interface, the efficiency and speed of the annotation process, and the seamless integration of new features into the data annotation platforms. An important aspect is to facilitate the annotators' tasks by empowering them with label suggestions for guidance. These suggestions provide valuable insights into the input data, enabling annotators to make more informed decisions during the annotation process. The main goal is to enhance the overall user experience by providing guidance through label suggestions, insights on the data, and seamless extensibility.

Label Suggestions: An auto-annotation modeling is crucial in an annotation tool as it facilitates the automation of data annotation through the application of machine learning techniques, thereby reducing manual labor and improving efficiency. The model's performance can be monitored by the expert human-in-the-loop, ensuring the accuracy and reliability of annotations. Additionally, auto-annotation mitigates human error, resulting in more consistent and reliable annotations. By incorporating this feature, the annotation tool can learn from previously annotated data and continuously improve its predictions, making it a valuable tool for large-scale annotation projects. Also, a user-friendly interface makes it easy for annotators to perform their tasks and reduces the risk of errors. An annotator-friendly user interface typically includes intuitive navigation, clear instructions, and simple tools for performing annotations. By providing a user-friendly interface, an annotation platform can improve the overall annotation process and ensure high-quality annotations. A user-friendly interface that is intuitive and easy to use also ensures that the tool is accessible to a broad range of users, thereby increasing its adoption and utility.

Data Insights and Annotator performance: Gaining insights into the amount of data required to achieve convergence yields valuable information about the quality and quantity of data needed for effective model training. The point at which the model's performance stops improving and reaches a plateau indicates that additional data will not significantly improve the model's accuracy. This information helps annotators determine the minimum amount of data required for effective training and optimize their data utilization and set the budget of annotation accordingly. By having insights into the percent of data used to attain convergence, annotators can balance the trade-off between the cost of acquiring additional data and the accuracy of the model. This information is essential for ensuring the quality and reliability of the annotations produced by the tool, and it enables annotators to make informed decisions about the data they use to train the model.

Seamless Extensibility: Seamless extensibility is important for an annotation platform because it allows new functionalities, models, and acquisition techniques in addition to the existing ones without causing disruption to the existing platform. This enhances the overall user experience and allows for customization to meet specific needs. With seamless extensibility, new tools, features, and integrations can be easily added as needed, making the platform more adaptable to changing requirements and enabling it to grow and evolve over time. Eventually, enabling organizations to continuously optimize and improve the tool, ensuring its relevance and usefulness over time.

4.2.3 Workflow Integration

Easy and effective integration of an auto annotation tool into an existing workflow or process for annotating data is very desirable. Integrating the tool with other tools and systems that are already being used in the annotation process, such as data annotation platforms, data management systems, and other software applications shows whether the system can scale or not. The goal of workflow integration is to streamline the annotation process and make it more efficient while ensuring that the annotations produced are of high quality and consistent with existing standards. Integrating the tool into the existing workflow becomes a seamless part of the process and reduces manual effort and time. It can include features like compatibility with multiple file formats, Scalability, and effective integration with existing workflows and tools.

Versatility: Versatility refers to the ability of the tool to have multifaceted file format support and the capability to handle a wide range of data types and formats. The ability to support multiple file formats and annotation types enables the tool to be used for a broader range of applications, as different data sources and annotation types require different methods for processing and storing data. This versatility also enables organizations to utilize the tool for a variety of projects, reducing the need for multiple specialized tools. The support for multiple file formats and annotation types also makes it easier to integrate the tool into existing workflows, as it can work with the data and systems that organizations are already using. The ability to handle a broad range of data types and formats is a crucial aspect of making the annotation tool useful and effective.

Scalibility Large-scale data handling is important in data annotation because it enables users to annotate vast amounts of data efficiently. The ability to handle large-scale data is critical for organizations to process large amounts of data, in the fields of natural language processing. Annotation tools that are unable to handle large-scale data may be inefficient and slow, requiring excessive amounts of time and resources to complete the annotation process. By contrast, tools that are designed to handle large-scale data can process data quickly and accurately, reducing the time and resources required for annotation tasks. The ability to handle large-scale data also enables organizations to annotate a greater volume of data, providing more comprehensive insights into their data and enabling more informed decision-making. Large-scale data handling is therefore a critical factor in the effectiveness and efficiency of data annotation tools.

Interoperability: Interoperability allows a platform to seamlessly integrate with existing workflows within an organization. This ensures that the annotation process can be smoothly incorporated into existing workflows, reducing the need for manual intervention and reducing the risk of errors. The integration of the annotation platform with existing workflows and tools also allows for the seamless exchange of data and annotations, enabling organizations to leverage their existing tools and data to improve the accuracy and efficiency of their annotations. Additionally, workflow integration interoperability helps organizations to reduce the time and resources required for annotation tasks, by automating repetitive tasks and reducing the need for manual data entry. The ability to integrate the annotation platform with existing workflows and tools also allows organizations to leverage their existing data and infrastructure, making it easier and more cost-effective to implement an annotation platform. The availability of APIs can potentially facilitate easy integration of the tool into existing data annotation platforms, thus enabling automation and streamlining of the annotation process. Free services, on the other hand, reduce the cost barrier for users and increase accessibility, thereby promoting the adoption and utilization of the tool.

4.3 Comparison of systems and tools

A comparison of different existing tools and systems has been presented in Table 4.1. The table provided offers a comprehensive comparison of different existing tools and systems based on various features crucial for efficient data annotation. For the comparison, we considered some common tools like Doccano (Nakayama et al., 2018), Label Studio (Tkachenko et al., 2020-2022), Light Tag (Perry, 2021), Prodigy (*Prodigy*, 2017), Label Sleuth (Shnarch et al., 2022), Potato (Pei et al., 2022), and Small-Text (Schröder et al., 2023). The systems that provide the given features

Table 4.1: Table of comparison for different existing tools and systems based on various features

Features	Systems	Doccano (Nakayama, Kubo, Kamura, Taniguchi, & Liang, 2018)	Label Studio (Tkachenko et al., 2020-2022)	Light Tag (Perry, 2021)	Prodigy (*Prodigy*, 2017)	Label Sleuth (Shnarch et al., 2022)	Potato Pei et al. (2022)	Small-Text (Schröder et al., 2023)
Machine Learning Capabilities	Active Learning		✓		✓	✓	✓	✓
	Few-shot Learning		✓		✓			✓
	Iterative fine-tuning	✓	✓	✓	✓	✓		✓
User Experience	Label Suggestions	✓	✓	✓		✓	✓	✓
	Data Insights & annotator performance	✓	✓	✓	✓			
	Seamless Extensibility	✓	✓	✓	✓	✓	✓	✓
Workflow Integration	Versatility	✓	✓				✓	
	Scalability	✓	✓					✓
	Interoperability		✓	✓				✓

are marked for those particular features and red marks signify that the feature exists but in a paid version for commercial use of the tool. Also, some of the features are not directly included in the tools but the system may provide a way for the user to integrate the feature. However, we will consider features that are directly provided by the systems in the table of comparison 4.1.

Among the mentioned systems, Doccano is a useful tool that possess several valuable features. It supports label suggestions, data and annotator performance insights, extensibility for adding new functionalities, and is scalable. However, it lacks certain key features like active learning for data sampling and Few-shot learning, which can enhance data annotations' efficiency. On the other hand, Label Studio[1] shows great potential and offers most of the desired features. It includes active learning capabilities, Few-shot learning, label suggestions, and performance insights, making it a robust platform for data annotation but it's important to note that these advanced features are only available in the paid enterprise version, which may require consideration based on specific project requirements and budget. Although, LightTag[2] offers an array of features in its free version, including iterative fine-tuning which allows progressive refinement of models for more accurate annotations, it only provides other desirable features such as label suggestions, insights on data and annotator performance, and seamless extensibility only in a paid version. Prodigy[3] is another annotation tool that provides active learning, few-shot learning and iterative fine-tuning. It also provides insights on data and annotator performance but it is not free or open-source to use. Among the listed features, Label Sleuth supports features like active learning, iterative fine-tuning, label suggestions, and extensibility but lacks many other key features. Also, it does not provide support for different text-based tasks. Similarly, Potato is a portable text annotation tool that allows active learning, label suggestions, extensibility and versatility but it does not provide a holistic tool inclusive of all properties. Other tools like Small-text exhibit Active Learning, Few-shot learning and iterative fine-tuning capabilities. It also provides label suggestions, extensibility, and subtle workflow integration but it only works for text classification tasks.

Considering the features presented in the table, it becomes evident that each system possesses a unique set of capabilities. However, the ideal annotation tool should encompass all these features to

[1] https://labelstud.io/guide/label_studio_compare.html
[2] https://www.lighttag.io/features
[3] https://prodi.gy/features/large-language-models

provide a holistic and powerful solution. The need for a comprehensive tool that can seamlessly integrate mentioned machine learning capabilities, user experience and workflow integration becomes apparent. Most of these tools like Label Studio are growing and progressing really fast and some of them provide a good deal of features in the form of an academic program where you can use extended features for non-commercial purposes. Overall, it was observed that the enterprise version of Label Studio possesses most of the properties and it is well-suited for various tasks. However, to offer users a more personalized experience and access to additional features like diverse active learning techniques and insights on data and user performance, we can develop a tool that provides flexibility, efficiency, and adaptability. By doing so, organizations can enhance their data annotation processes and achieve superior model performance while optimizing resource utilization.

4.4 Summary

To summarize this chapter, we provide the reader with a comprehensive comparative framework, evaluating existing text annotation tools based on specific criteria essential for a robust text annotation platform. By carefully examining the essential features required in a text annotation system, we provided valuable insights into why these functionalities are required and how they contribute to building a powerful and reliable system. Through a critical analysis of the strengths and weaknesses of the current tools in the field, we identified potential areas of improvement and highlighted the challenges faced in the annotation process. By harnessing the collective power of these functionalities, we aim to provide a direction towards a holistic system for text annotation which would not only address the limitations of existing tools but also streamline the annotation workflow, saving valuable time and effort for researchers catering to diverse user needs. In conclusion, this chapter sets the stage for the development of a state-of-the-art text annotation platform that meets the dynamic demands of modern natural language applications. By leveraging the insights gained from our comparative analysis, we can offer users a powerful platform for text annotations, addressing the shortcomings of existing tools and enhancing the overall annotation process.

Chapter 5

Towards efficient text annotation: System Architecture and Evaluation

5.1 Overview

In the previous chapter, we talked about a comprehensive framework including the existing text annotation systems and their key features. We will now move towards an efficient approach to text annotation leveraging various trending concepts. In this chapter, we will discuss the system architecture of the approach to an efficient text annotation platform, key components of the system, its design, and the development process that was followed. Text annotation plays a vital role in various natural language processing tasks, enabling machines to understand and interpret human language effectively. Therefore, we will also explore various different natural language-based tasks and use cases validating the system's power and versatility. Also, we will evaluate the system performance and show the adequacy of concepts we used in the system like active learning and few-shot learning. We will present how it has been instrumental in various tasks like the task of Question Understanding in a question-answering system which will be discussed in detail in the coming sections, and other tasks like Named Entity Recognition, Sentiment Analysis, Question Classification, etc. The main aim of this chapter is to provide insights into how we can improve the process of text annotation for enabling more sophisticated AI systems.

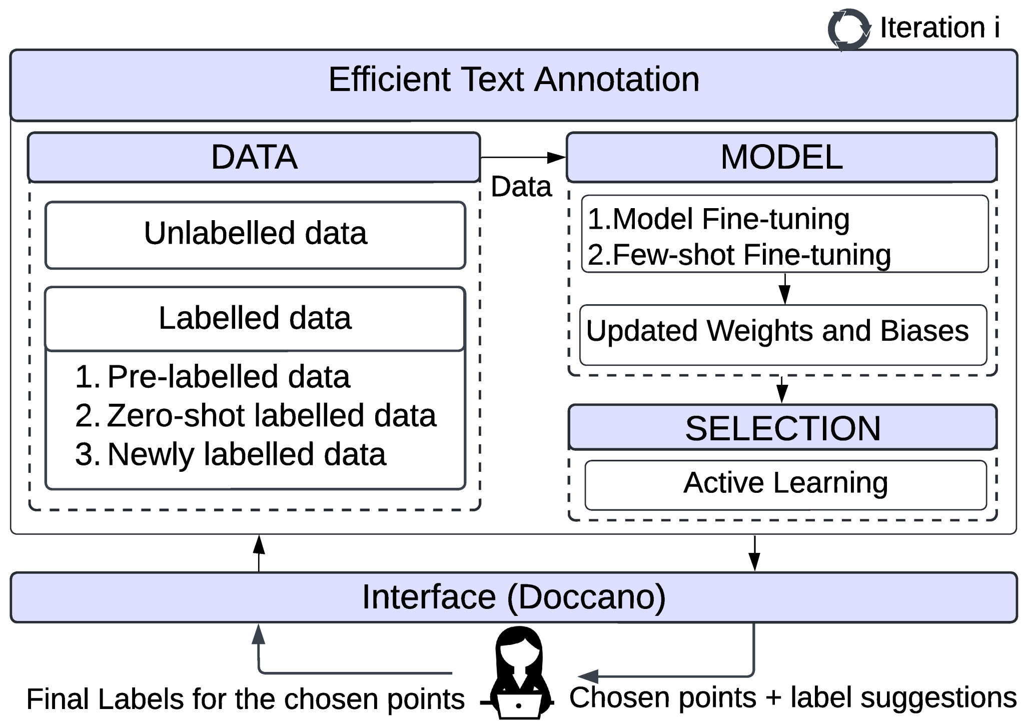

Figure 5.1: Iterative model training using active learning and few-shot learning for optimal performance enhancement

5.2 Efficient Text Annotation: System Architecture

The system architecture for the efficient text annotation platform comprises of three main modules for data handling, acquisition, and model training and inference. The architecture of the system as shown in Figure 5.1 describes the components of each module. The data handling module performs data pre-processing such as tokenization, stemming, and stop word removal depending on the task and dataset. It enables the system to accept multiple tasks and their corresponding data, allowing users to work with different text annotation tasks such as Named Entity Recognition, Sentiment Analysis, and Topic Classification. The second module of the system deals with the data selection strategies for the acquisition of the best data points as per the model. It provides the user with active learning support for the system incorporating various active learning and deep active learning techniques. It provides an on-demand tool for active learning, where the model is initially trained on a small set of labeled data, and then the acquisition module is used to select the most informative data points for further annotation. After annotating the selected data points, they are used to fine-tune the model using a few-shot learning technique. The third module in this system architecture deals with model training and model inference. It supports various language models that can be fine-tuned as per the user's need. The system also utilizes pre-trained models which are trained on a large corpus of unsupervised text data and can be fine-tuned on the user's data, while other models are trained from scratch on the user's data. The system also supports a Few-shot fine-tuning mode where we can perform few-shot fine-tuning of sentence transformers by changing a few parameters in the system settings. This module enables the system to support different types of model architectures such as CNN, CNN_BiLSTM_CRF, and models like SetFit, and BERT. Moreover, it has the ability to work with a human-in-the-loop approach, where the model's predictions are reviewed and corrected by human annotators, before being sent to the model to be fine-tuned to improve the quality and accuracy of the predictions in the next iteration.

As shown the Figure 5.1, the user's first touch point is via a nice interface that is inspired by the open-source tool for annotations called Doccano (Nakayama et al., 2018). As the user gives an input to the system, it is directly taken to the data module where it pre-processes the data to make it ready for further steps. Its accepts input in multiple different formats. Depending on the amount of data that is initially labeled, the system keeps a record of data in the form of two sets, labeled and unlabeled. The labeled set could be retrieved from the pre-provided set of labeled data by the user, or if the user has no labeled data to begin with then we can use a zero-shot learning-based extension of the system that will ask the user for a template-based prompt related to the underlying task and data, and get the initial set ready. It will also take care of the data that the human-in-the-loop labels in an iteration updating the two sets of data before model fine-tuning in every iteration. We further pass the data to the next module which deals with model fine-tuning and inference. In this module, the labeled set of data is utilized for the process of model fine-tuning, and the unlabeled set is used to make predictions for the data selection and labeling process. The user can choose the model they want to use and they can also choose whether they want to perform standard fine-tuning or a few-shot learning-based fine-tuning. After fine-tuning, the updated model is used to generate predictions on the unlabeled set of data. These predictions are further given to the active learning data selection module which ranks the data points based on various active learning techniques and selects a set of points to be labeled by the human expert. The user can choose the active learning technique that they want to deploy and the selected data points are presented back to the oracle who is the human-in-the-loop annotating the given data. The oracle receives this set of data points along with label suggestions and they can decide if the label suggestions are good or if they need any changes. The platform allows the oracle to make the required changes and move further with this process of annotating data in iterations. The updated labeled data is then fed back into the system and the cycle continues, leading to an accurate model with a reduced data labeling cost.

5.3 System Design and Key Concepts

To embark on this approach, it is crucial to develop a text annotation platform that optimizes the annotation process and maximizes the utilization of labeled data. In this section, we present the design of the efficient text annotation approach. This approach incorporates three key concepts: Active learning-based sample selection, Few-shot learning, and guided annotations with iterative model training based on the response from the human-in-the-loop. We will also discuss the significance of each of these aspects and their role. Let's discuss each of these concepts in detail:

(1) **Active Learning**: Active learning techniques play a vital role in reducing labeling efforts by selectively choosing instances for annotation that are most informative for model improvement. Instead of annotating all the unlabeled instances, active learning comes into play by selecting a subset of instances that the model finds uncertain or challenging. The process of active learning starts with an initial set of labeled data, which is used to train an initial model. The trained model is then used to make predictions on the remaining unlabeled pool of data which are analyzed, ranked, and selected based on the active learning techniques. By focusing on these instances, active learning aims to acquire new annotations strategically, maximizing the impact of each annotation on model performance. Annotating the instances that are most informative or difficult for the model to understand can help improve its accuracy and generalization capabilities. The integration of active learning within our text annotation platform brings significant benefits. Firstly, it reduces the annotation burden by significantly reducing

the number of instances that need to be labeled. Instead of annotating the entire dataset, annotators can focus their efforts on the instances that are most valuable for improving the model's performance. Secondly, active learning enables a feedback loop between the annotators and the model. As annotators label new instances, the updated labeled data is incorporated into the training process, allowing the model to further refine its predictions. This iterative process of annotation and model improvement can lead to a virtuous cycle, where the model becomes more accurate with each round of annotation. The platform is designed in such a way that the user can move between data points to be labeled and these data points are always presented to the user in an active learning-based order. The user is always presented with the next best data point as they go further with the annotation process. The user can choose the active learning technique they want to use and some of the selection techniques that are supported by the system are the Uncertainty approach, Least Confidence, Core-set approach, Margin Sampling, Maximum Normalized Log Probabilities (Shen et al., 2017), Deep Bayesian active learning, and random selection. Some of the challenges of this approach were regarding the adaptation of new labels in between the annotation process and the cold-start problem in case the user has a small or no set of data, to begin with. These challenges brought us to choose our next important aspect of the system which is few-shot learning.

(2) **Few-Shot Learning**: Few-shot learning allows models to generalize from a small number of labeled examples to perform well on the given tasks. The text annotation platform leverages Few-shot learning models to provide model adaptability to newly added information and to address the cold-start problem. It also helped in the guidance for the annotation process which essentially provides annotation suggestions to the oracle to maximize model performance with minimal labeling efforts. One approach to utilizing few-shot learning models in guiding the AL process involves fine-tuning a Few-shot learning model like SetFit (Tunstall et al., 2022) on a small set of labeled data. This training captures task-specific knowledge and enables the model to adapt quickly to new tasks with a small or no initial set of labels. Using the few-shot learning model, the platform identifies instances that are more likely to benefit the active learning process. This approach effectively selects instances that maximize the

model's performance with minimal labeling efforts. Another approach to address the common problem of cold start is to leverage zero-shot learning techniques (Sanh et al., 2022). Utilizing zero-shot learning can address the challenge of cold start by enabling the model to make predictions on unseen classes without prior training. It enables the model to make predictions on classes or tasks it has not been explicitly trained on. This approach leverages pre-existing knowledge and semantic relationships to generalize to new categories. By incorporating zero-shot learning, the system becomes more adaptable and capable of handling new and unseen data. So, the system is designed in such a way that even if we start with a very small amount of data, the model will be able to adapt.

(3) **Guided annotations with iterative model training**: The proposed technique uses suggestions from the model that is fine-tuned over iterations. Annotating large data sets can be challenging and more prone to errors and bias. However, following this technique of providing label suggestions can be really useful for training accurate and robust models. The platform aims to streamline the annotation process, providing intuitive tools and workflows that enable annotators to annotate large volumes of data efficiently. Model training and guided annotations are integral to the efficient annotation process. The platform incorporates model training to develop accurate and robust models that can perform complex text-based tasks. By training models on labeled data, the platform empowers the models with task-specific knowledge and enhances their ability to understand and interpret the text. The facilitation of guided annotations not only helps annotators provide context-specific guidance and suggestions during the annotation process but also enables a cycle of human feedback for tricky scenarios like dealing with newly added labels. This guidance comes from pre-trained models, which are fine-tuned and improved over iterations. Guided annotations help annotators work more efficiently and consistently by leveraging the knowledge encoded in the models. The system by design provides the user with a choice to let the user enable this auto-labeling mode if they want. The suggestions can be received from a pre-trained model with standard fine-tuning or few-shot learning-based fine-tuning. The user can also choose the model and the settings they want to utilize in order to get these suggestions and start this iterative process.

To summarize, we adopt an iterative training strategy combining active learning and few-shot learning, leveraging the model's confidence and performance gain to select instances for labeling and then using this labeled set for fine-tuning the models. This iterative process progressively improves the model's performance while minimizing labeling efforts and computational requirements. It allows the human expert in the loop to make decisions based on the suggested labels by the system, and correct the existing labels accordingly. By combining these concepts, the platform optimizes the annotation process by incorporating a streamlined annotation workflow increasing the annotation consistency with efficient resource allocation, and maximized utilization of labeled data.

5.4 Evaluations

In this section, we will discuss various use cases that demonstrate the effectiveness and versatility of our text annotation approach. By leveraging this methodology, we can extract meaningful information from textual data. Through the process of annotating text, we empower machine learning models to understand and interpret unstructured textual information more accurately. The use cases presented here show a wide range of applications where text annotation plays a pivotal role. We will look into diverse domains such as sentiment analysis, named entity recognition, text classification, question-answering types, and text generation. Each of these use cases highlights the significance of utilizing a text annotation approach to enhance and optimize the learning process. By examining these use cases, we aim to shed light on the potential of text annotation in addressing real-world challenges and providing solutions across different industries. Moreover, we unravel the correlation between active learning and few-shot learning techniques, which further amplifies the effectiveness of the annotation process. In the subsequent sections, we will explore each use case in detail, studying various tasks and datasets we used for these tasks 7.1. We will also discuss various annotation strategies and their performance. By examining these use cases, we hope to provide a comprehensive understanding of the wide-ranging capabilities and advantages that this text annotation approach offers.

In the field of text classification, the role of a high-quality text annotation platform cannot be overstated when it comes to training accurate and reliable machine learning models. The various tasks that we will present include sentiment analysis using customer review data (X. Ding, Liu, & Yu, 2008), adverse drug reaction classification (Gurulingappa et al., 2012), expected answer prediction (Kwok, Grunfeld, Dinstl, & Chan, 2000), Toxicity classification using Toxic Conversation data[1] where toxicity is defined as anything rude, disrespectful or otherwise, spam detection using Enron dataset (Metsis, Androutsopoulos, & Paliouras, 2006), etc. By leveraging an efficient text annotation platform, we were able to streamline the annotation process, resulting in faster and more accurate labeling of large datasets. The platform allowed us to define annotation guidelines, create annotation tasks, and manage the annotation process effectively. In our experiments, we explored the benefits of incorporating active learning and few-shot learning techniques within these text classification tasks. Active learning, for instance, played a significant role in our experiments. By actively selecting the most informative samples for annotation, we were able to reduce annotation efforts while maintaining high-quality training data. This iterative process of selecting the most valuable samples for annotation allowed us to optimize the annotation process and achieve better results. We initiated our experiments by utilizing the efficient text annotation platform for various text classification tasks, starting with sentiment analysis. Additionally, we applied the platform to tasks such as adverse drug reaction classification, expected answer prediction, multi-class text classification, and document classification. We also utilize the datasets employed in setfit (Tunstall et al., 2022) to conduct experiments on Enron spam detection, Twitter complaints, etc. Using the platform, a diverse range of texts can be annotated including product reviews, social media posts, and customer feedback, to build a sentiment analysis model capable of accurately gauging sentiment across different domains.

In addition to the task of text classification, the text annotation platform also proved valuable in Sequence Labelling tasks like named entity recognition. Named entity recognition focuses on identifying and classifying entities within a given text, including names of people, organizations, locations, and more. Leveraging the annotation capabilities of the platform, we were able to annotate a significant dataset to train a named entity recognition model. The process involved annotating the dataset with specific labels to mark the boundaries and types of named entities present in the text. This annotation process, enabled by the text annotation platform, played a crucial role in generating

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

high-quality training data for the named entity recognition model. The annotated dataset served as a foundation for training the model to accurately identify and classify named entities in new texts. By leveraging the labeled data, the model could learn patterns and characteristics associated with different types of named entities, thereby improving its ability to accurately identify and classify entities in unseen text. One of the ways to Few-shot Learning can be integrated with active learning for the sequence labeling task of recognizing and classifying named entities in text using only a few labeled examples per entity class could be using NNShot and StructShot techniques (Yang & Katiyar, 2020a). NNShot is a simple token-level nearest neighbor classification system that computes a similarity score between a token in a test example and all tokens in a support set, and assigns the token a tag corresponding to the most similar token in the support set. STRUCTSHOT is an extension of NNShot that models label dependencies with a simple Viterbi decoder. This could particularly be useful in scenarios where new entity types emerge or when there is a need for immediate support for emerging entity types without retraining and redeploying the named entity model.

Let's explore six different use cases for text annotation, each highlighting a specific task and its practical applications as shown in Table 7.1. For each use case, we will provide an overview of the task, mention data set used with the number of queries and describe the experimental settings with results. Furthermore, we will discuss the results and compare the performance of the active learning based normal fine-tuning with experiments incorporating Few-shot learning-based fine-tuning in each iteration, which will allow us to maximize the efficiency of utilizing limited labeled data.

Table 5.1: Table presenting different tasks and data sets in text annotation used in experiments

	Task	**Dataset**	**Size**	**Annotation Type**
1.	Question Understanding	LCQuAD, QALD-9	1700	Text Generation
2.	Question Classification	TrecQA	5.5k	Multi-class
3.	Named Entity Recognition	CONLL-03	14k	Sequence Labelling
4.	Sentiment Analysis	Customer Review	30k	Binary Classification
5.	Spam Detection	Enron Spam	33k	Binary Classification
6.	Toxicity Classification	Toxic Conversation	1.75M	Binary Classification

5.4.1 Question Understanding in KGQAn system

Let us discuss our first use case where we used the approach mentioned above as the underlying concept in the context of text generation. This use case was part of a research project called KGQAn (Omar, Dhall, Kalnis, & Mansour, 2023) which is "A Universal Question-Answering Platform for Knowledge Graphs" that we published in Proceedings of the ACM on Management of Data and the 2023 ACM SIGMOD/PODS Conference. The proposed text annotation approach plays a crucial role in this use case for annotating over 1700 questions for question understanding in KGQAn. KGQAn proposed a novel approach that eliminates the need for tailoring the question-answering system to each individual Knowledge Graph. It acts as an intermediary between users and any available Knowledge Graph, and it consists of three main modules of question understanding, just-in-time linking, and answer-type filtering. The question understanding approach in KGQAn is structured as a text generation task, extracting abstract triple patterns from natural language questions. In the question understanding approach of KGQAn, we undergo a comprehensive training process using a sequence-to-sequence (Seq2Seq) (Sutskever, Vinyals, & Le, 2014) model, exposing it to a diverse range of questions. This training ensures that KGQAn becomes adept at comprehending various question types and generating accurate responses. Generating coherent and contextually relevant text required a well-annotated dataset that covered a wide range of topics, styles, and genres. The proof-of-concept text annotation platform facilitated efficient annotation of a large corpus of text, enabling the creation of a diverse training data set. This diverse training data significantly enhanced the language model's ability to generate high-quality annotations. The annotation process employed in KGQAn is highly tailored to this individual use case and hence necessitates expert knowledge. The formalization of question understanding as a text-generation task and the meticulous annotation process using the proposed system ensured KGQAn's accurate delivery of results and valuable responses to a wide range of user queries. To further understand and read how each module affected the question-answering technique followed in KGQAn, please refer to this published work (Omar et al., 2023)

To illustrate the annotation process in more detail, let's consider an example from the actual annotated dataset. The questions used in KGQAn were sourced from two well-established benchmark

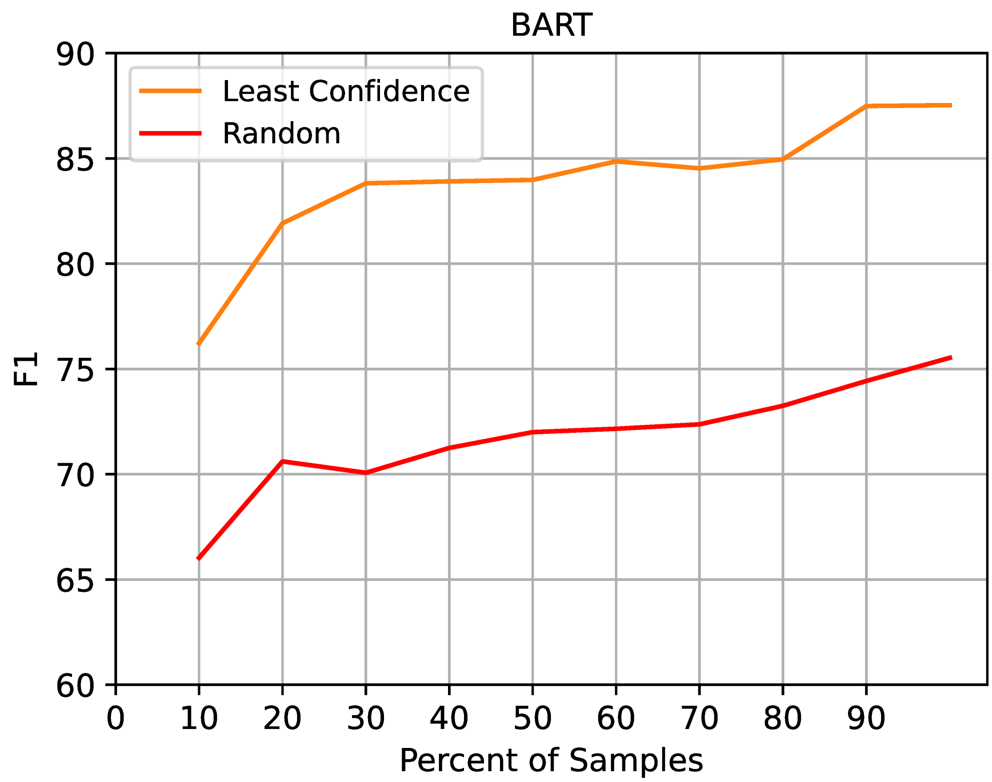

Figure 5.2: Iterative model training of Bart on use-case specific dataset for the task of Question Understanding for Knowledge Graphs

datasets known as LC-QuAD-1.0 (Trivedi, Maheshwari, Dubey, & Lehmann, 2017) and QALD-9 (Usbeck, Gusmita, Ngomo, & Saleem, 2018). Let's examine a specific question from the annotated set: "Name the origin of Henry Cluney." The annotation process involves identifying and representing the relevant information in the form of triples, which abstractly represent the given text as a graph. These annotations are stored in JSON files for further use. Our annotation process adheres to a systematic strategy that involves several steps. Firstly, we extract the entities and relations present in the question. Next, we classify the entities into two categories of 'named entities' and 'variables.' This categorization helps in differentiating between specific entities and the placeholders for the unknowns in the question. For the named entities, we provide essential information such as the entity value itself representing a concrete, identifiable entity mentioned in the question. For the variables, we go a step further and provide additional details. These include the expected answer type for the variable, such as a string or numeric value, semantic class types to indicate the general domain or category to which the answer belongs (e.g., place, person, organization) and a variable ID that serves as a unique identifier for each variable in a given query. This variable ID plays a crucial role in identifying and tracking the unknowns throughout the annotation process. By incorporating these steps and providing comprehensive information for the unknowns, we ensure that the annotation captures all necessary details for effective question answering. This structured annotation approach enhances the KGQAn system's ability to understand the question and generate accurate responses, even for queries involving multiple unknown variables. After following the steps mentioned, we receive the annotation for the mentioned question as follows:


```
{"id": "794",
 "text": "Name the origin of Henry Cluney ?",
 "triples": [{
    "Entity1": {
        "value": "Henry Cluney",
        "type": "named entity" },
    "Entity2": {
        "value": "Name",
        "type": "variable",
        "ans_type": "string",
        "class": "place",
        "var_id": 1 },
    "Predicate": "origin"}]
}
```


In this annotation, we have Entity1 which represents the named entity "Henry Cluney" in the question, Entity2 which represents the variable "Name" referring to the unknown whose answer needs to be retrieved, and a predicate that represents the relationship or property being queried, "origin". Entity2 has additional information such as answer type (a string in this case), class (indicating the semantic type which is place in this case), and variable id (providing a unique identifier). By annotating the question with triples and storing the resulting annotations in JSON files, we create a structured representation that captures the essential information needed for the KGQAn system. These annotations serve as valuable training data for the question-understanding module of the system, enabling it to understand and generate accurate responses to similar questions. We use parts of these annotations and fine-tune the Sequence2Sequence model.

In the experiment, we used the BART model (M. Lewis et al., 2019) with a batch size of 4 for 5 epochs while sampling 10% of data in every iteration for 10 iterations. The results of the experiment are presented in Figure 5.2 which are the average of five runs. We sampled data points based on the model's confidence using an uncertainty-based active learning technique. The uncertainty active learning technique played a pivotal role in this annotation process as the annotation effort required was significantly reduced while maintaining the quality of the training data. It allowed for the iterative selection of the most informative samples for annotation. We were able to perform better using the similar amount of data as shown in the Figure.

5.4.2 Question Classification

The task of Question Classification is to assign a specific category or class label to a given question. The main goal of this task is to automatically classify questions into predefined categories, enabling efficient organization and retrieval of information. Question Classification can be really useful in various applications, such as information retrieval systems, question-answering systems, and chatbots, as it helps in understanding the user's query and providing relevant responses or information. The Text REtrieval Conference (TREC) Question Classification dataset is a widely used data set for the task of question classification. It comprises a training set with 5500 labeled questions and a test set containing 500 questions. The dataset includes 6 coarse class labels and 50 fine class labels, enabling fine-grained categorization of the questions. On average, each sentence in the dataset has a length of 10 words, and the vocabulary size is 8700. We used a CNN model and 5 different active learning techniques in the experiment. All values in Figure 5.3 are the average of five runs. This configuration achieved convergence at 45% of the training data using the Bayesian Active Learning by Disagreement with Monte Carlo Dropout (BALD). The experiment shows the effect of the active learning process in successfully reducing the amount of labeled data required to achieve convergence.

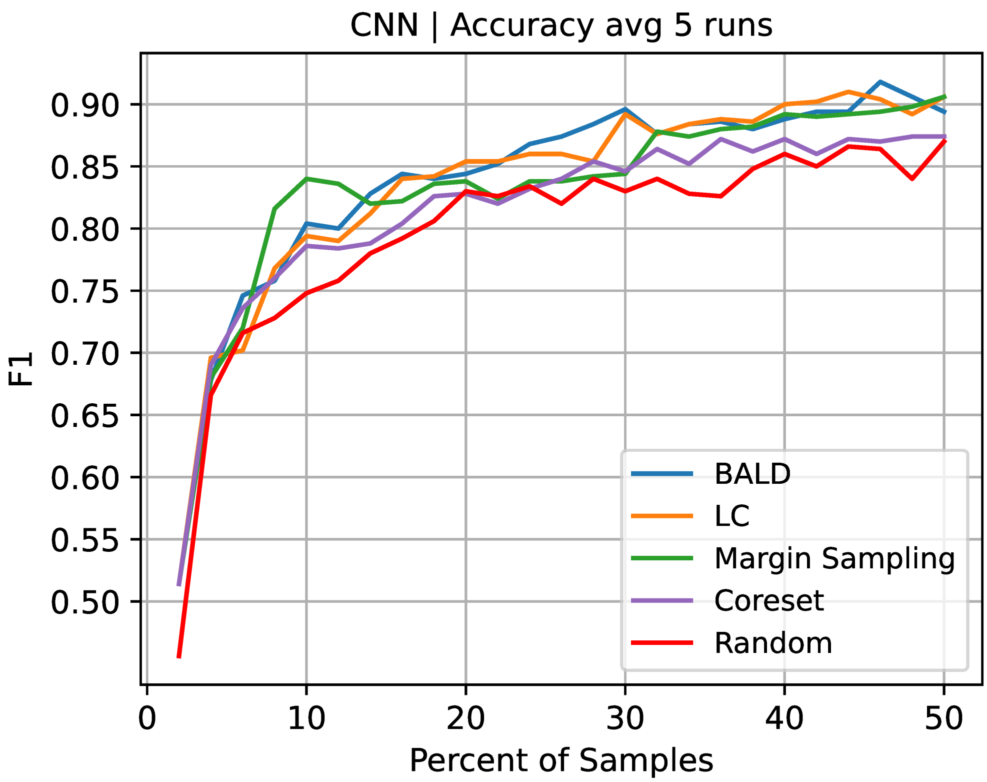

Figure 5.3: Iterative model training of CNN on TRECQA dataset for the Question Classification

5.4.3 Named entity Recognition

Named Entity Recognition (NER) is a natural language processing task that involves identifying and classifying named entities in text. The main goal of this task is to accurately recognize and label the given entities in text. This task is crucial for various applications, including information extraction, question answering, and text summarization. NER algorithms typically use machine learning or deep learning techniques to analyze the contextual information surrounding words and determine their entity types. The output of NER is a sequence of labeled entities, indicating their positions and corresponding entity categories in the text. By accurately identifying named entities, NER enables better understanding, organization, and analysis of unstructured textual data. CONLL 2003 (Sang & Meulder, 2003) is a benchmark dataset that is very commonly used for NER, consisting of eight files for English and German. It was released as part of a shared task for language-independent NER

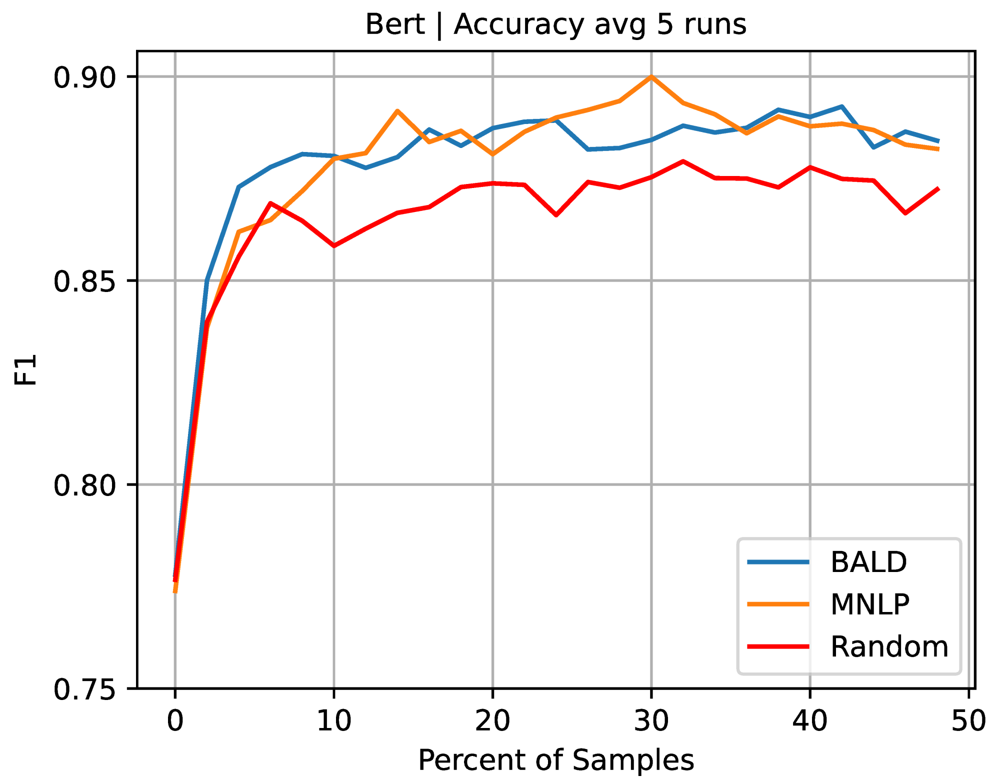

Figure 5.4: Iterative model training Bert-base-uncased model on CONLL-03 dataset for Named Entity Recognition

and is based on the Reuters Corpus for English, with training and development sets taken from August 1996 and test sets from December 1996. The evaluation metric used in all experiments is the F1 score. For the named entity recognition task on the CoNLL 2003 dataset, we experimented using a pre-trained BERT-based uncased model. All values in Figure 5.4 are the average of five runs. The configuration achieved convergence in the 15th iteration of the cycle using 30% of the training data. The experiment shows the effect of the active learning process in reducing the amount of labeled data required to achieve convergence.

5.4.4 Sentiment Analysis

The goal of a Sentiment Analysis task is to analyze the customer reviews of products. A given query of a customer's review can be classified into positive or a negative class. The goal of the task is to work on the problem of determining the semantic orientations of opinions expressed on

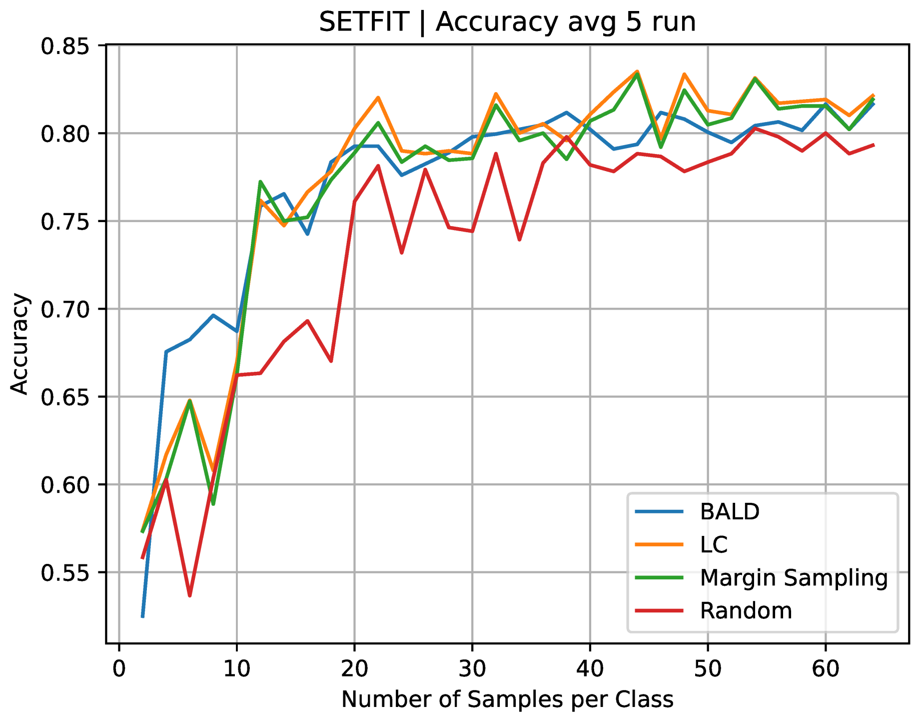

Figure 5.5: Iterative model training SetFit transformer model on Customer Review dataset for Sentiment Analysis

product features in reviews. The given task is essential in various applications, such as opinion mining, summarization, and search. The data set includes 3.39k sentences in the training data and 376 sentences in the test data. All values in Figure 5.5 are the average of five runs. In the experiment shown in the figure, we use the SetFit model for few-shot learning-based fine-tuning. This configuration achieved convergence the fastest in the case of Least Confidence and Margin Sampling. This shows that deep active learning techniques like Bayesian active learning are not always the best techniques for data selection and this can vary depending on the task and the data. Also, we reached the model convergence much faster in this case when compared to our previous tasks as we use a few-shot learning based fine-tuning.

5.4.5 Spam Detection

We also evaluated our proof-of-concept system on the Enron Spam Email data set for spam detection (Metsis et al., 2006). The task of spam detection involves classifying emails as either

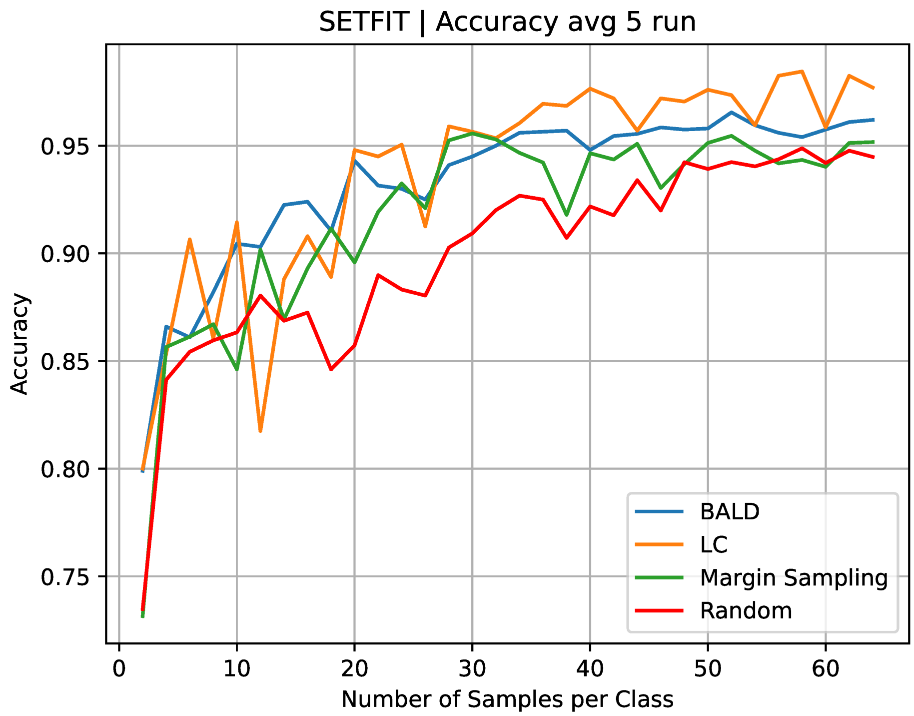

Figure 5.6: Iterative model training SetFit transformer model on Enron dataset for Spam Detection

”spam” or ”ham” (non-spam) where the main objective is to determine whether a given email, comprising the message content, should be labeled as spam or not. The data set comprises of 31k queries in the training set and 2k queries in the test set. This task is generally useful in scenarios where effective email filters and anti-spam systems that help users identify and manage spam emails, thus enhancing email security and user experience are needed. The results shown in Figure 5.6 are an average of five runs and uses the SetFit model for few-shot learning-based fine-tuning. As per the results, the least confidence based active learning sampling technique shows the best results and BALD also shows comparable performance when compared to other shown techniques.

5.4.6 Toxicity Classification

The toxicity classification task aims to classify the given text into two classes namely, toxic and non-toxic. We used the Jigsaw Unintended Bias in Toxicity Classification data set which contains comments from the Civil Comments platform together with annotations if the comment is toxic or

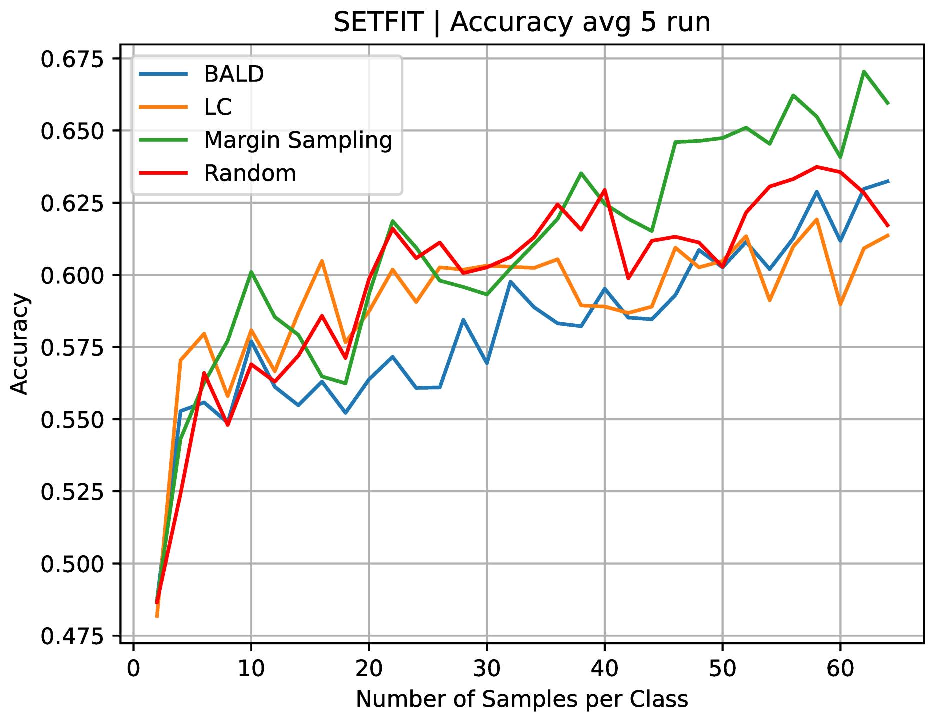

Figure 5.7: Iterative model training SetFit transformer model on Toxic Conversation dataset for Toxicity Classification

not. The data set is imbalanced, with only about 8% of the comments marked as toxic. We used the data set from the hugging face APIs [2]. All values in Figure 5.7 are an average of five runs. The configuration achieved convergence much earlier using the Margin Sampling active learning technique. The experiment uses both functionalities of the system, active learning based sampling and few-shot based fine-tuning and three different strategies as shown in the figure.

5.5 Summary

In this chapter, the focus shifts towards developing an efficient approach to text annotation by leveraging trending concepts. The groundwork laid in the previous chapter included an extensive framework covering existing text annotation systems and their key features. Building on that foundation, the chapter explores the system architecture of the proposed efficient text annotation

[2] https://huggingface.co/datasets/SetFit/toxic_conversations

platform. Key components of the system, its design, and the development process are thoroughly discussed. We explore various natural language-based tasks and use cases to validate the system's power and versatility. These tasks include Question Understanding in a question-answering system, Named Entity Recognition, Sentiment Analysis, and Question Classification, among others. In conclusion, the techniques discussed in this study can be applied effectively to a wide range of tasks and data distributions. Specifically, the SetFit experiment, which leverages few-shot learning-based transformer fine-tuning, demonstrates the benefits of combining few-shot learning with active learning when dealing with a limited initial dataset. Additionally, it is important to note that there is no one-size-fits-all active learning technique that outperforms all others; the effectiveness of each technique varies depending on the nature of the task at hand. Therefore, it is crucial to carefully choose the active learning strategy based on the specific characteristics and requirements of the problem being addressed. The main goal of this chapter is to provide valuable insights into enhancing the text annotation process. By streamlining and improving this crucial aspect, the groundwork is laid for the development of more advanced and capable AI systems that can effectively process and understand natural language data. Through the development of a text annotation platform for active learning and few-shot learning, we were able to address various annotation needs for tasks like question answering, text generation, named entity recognition, sentiment analysis, etc. By incorporating active learning techniques, we achieved efficient annotation of large datasets while maintaining high-quality training data. The versatility of the platform opens doors to countless other use cases and paves the way for future work.

Chapter 6

Conclusion and Future Work

In this thesis, we learned that the process of text annotation can be formidable and is quite critical for many natural language processing tasks. Our exploration presents various features desirable for a text annotation platform in the form of a comprehensive comparative framework over various existing tools. This exploration of various techniques and existing tools also revealed that there are many shortcomings in the existing techniques, and there is a lack of a holistic platform that can be used to mitigate the annotation challenge. To address this gap, we proposed a comprehensive approach that consists of three main components: Active Learning based efficient text selection, Few-shot Learning, and assisted annotations with iterative model training. By integrating these elements, we hope to empower researchers and practitioners to tackle the challenges of text annotation efficiently and effectively. To demonstrate the practicality and effectiveness of our annotation approach, we tried using it for a task that was part of a Question Answering platform, which was subsequently published in SIGMOD record 2023 (Omar et al., 2023). The results obtained from this real-world use case further validate the utility and potential of this annotation approach in addressing complex and challenging natural language processing tasks. As we continue to refine and expand our approach, we envision its broader application across a wide range of natural language processing domains, contributing to the advancement of text annotation methodologies and enriching the landscape of AI-driven applications. This motivated us to use the approach further to multiple other natural language processing tasks across diverse domains. Through these experiments, this research work aimed to provide the reader with insights into the approach's adaptability and robustness.

Nevertheless, text annotation remains a complex and ever-evolving field, and further research and improvements are warranted to continue advancing its capabilities. As the future of text annotation is envisioned, we envisage the development of unified platforms that would seamlessly integrate various trending concepts like active learning and few-shot learning. Such platforms would serve as powerful resources, empowering researchers to unlock the full potential of text annotation.

There are several exciting avenues for future research and development in this research work. Let's outline some key directions that can further enhance the capabilities and applications for text annotation in terms of the different concepts we saw in this thesis. One interesting direction could be to incorporate Data Shapely values (Ghorbani & Zou, 2019) into the text annotation process as it could potentially offer a comprehensive understanding of the individual data point importance in shaping the predictions of the model. The annotations' prioritization and resource reallocation can be improved by focusing on the most influential data points and quantifying each data point's impact. One such collaboration was successfully presented in an Active Data Shapely work (Ghorbani, Zou, & Esteva, 2022) which introduces a filtering layer in batch active learning by pre-selecting the highest-value points from an unlabeled dataset. Another good direction for the nature of the task in hand could be to harness reinforcement learning with a human-in-the-loop. It has tremendous potential as it allows an interactive loop between human annotators and the model, enabling active collaboration and continuous improvement. This work can further be expanded by making it compatible with existing novel tool kits like speech brain (Ravanelli et al., 2021) making it a multi-modal approach for tasks involving Speech, Images, Videos, and other kinds of data as it would unlock new opportunities for understanding complex real-world scenarios.

In summary, this work aims to contribute to the evolving field of text annotation and is committed to further bringing innovation in this domain. As the field progresses, the emergence of novel techniques and comprehensive platforms will propel text annotation to new heights, making a lasting impact in the field of natural language processing and beyond.

Chapter 7

Appendix

7.1 Courses

Course	Course code	Semester	Grade
DISTRIBUTED SYSTEM DESIGN	COMP-6231	Winter 2021	A
ADV. PROG. PRACTICES	SOEN-6441	Winter 2021	A-
FOUNDATIONS/SEMANTIC WEB	COMP-6531	Fall 2021	A+
BIG DATA ANALYTICS	SOEN-6111	Winter 2022	A+

7.2 Publications

7.2.1 Journal and Conferences

Reham Omar, **Ishika Dhall**, Panos Kalnis, **Essam Mansour**: A Universal Question-Answering Platform for Knowledge Graphs. Proc. of the ACM on Management of Data, V1 (**SIGMOD 2023**)

7.2.2 Demos

Reham Omar, **Ishika Dhall**, Nadia Sheikh, **Essam Mansour**. A Knowledge Graph Question-Answering Platform Trained Independently of the Graph. International Semantic Web Conference (**ISWC 2021**). https://ceur-ws.org/Vol-2980/paper312.pdf

References

Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., & et al. (2021). RAFT: A real-world few-shot text classification benchmark. In *Proceedings of the neural information processing systems track on datasets and benchmarks 1, neurips datasets and benchmarks 2021, december 2021, virtual.* Retrieved from https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/ca46c1b9512a7a8315fa3c5a946e8265-Abstract-round2.html

An, B., Wu, W., & Han, H. (2018). Deep active learning for text classification. In *Proceedings of the 2nd international conference on vision, image and signal processing, ICVISP 2018, las vegas, nv, usa, august 27-29, 2018* (pp. 22:1–22:6). ACM. Retrieved from https://doi.org/10.1145/3271553.3271578 doi: 10.1145/3271553.3271578

Asghar, N., Poupart, P., Jiang, X., & Li, H. (2017). Deep active learning for dialogue generation. In N. Ide, A. Herbelot, & L. Màrquez (Eds.), *Proceedings of the 6th joint conference on lexical and computational semantics, *sem @acm 2017, vancouver, canada, august 3-4, 2017* (pp. 78–83). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/S17-1008 doi: 10.18653/v1/S17-1008

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., & Agarwal, A. (2020). Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, april 26-30, 2020.* OpenReview.net. Retrieved from https://openreview.net/forum?id=ryghZJBKPS

Atighehchian, P., Branchaud-Charron, F., & Lacoste, A. (2020). Bayesian active learning for production, a systematic study and a reusable library. *CoRR*, *abs/2006.09916*. Retrieved from https://arxiv.org/abs/2006.09916

Ayub, A., & Fendley, C. (2022). Few-shot continual active learning by a robot. In *Neurips*. Retrieved from http://papers.nips.cc/paper_files/paper/2022/hash/c58437945392cec01e0c75ff6cef901a-Abstract-Conference.html

Azimi, J., Fern, A., Fern, X. Z., Borradaile, G., & Heeringa, B. (2012). Batch active learning via coordinated matching. In *Proceedings of the 29th international conference on machine learning, ICML 2012, edinburgh, scotland, uk, june 26 - july 1, 2012.* icml.cc / Omnipress. Retrieved from http://icml.cc/2012/papers/607.pdf

Bart, E., & Ullman, S. (2005). Cross-generalization: Learning novel classes from a single example by feature replacement. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005), 20-26 june 2005, san diego, ca, USA* (pp. 672–679). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2005.117 doi: 10.1109/CVPR.2005.117

Bennequin, E., Bouvier, V., Tami, M., Toubhans, A., & Hudelot, C. (2021). Bridging few-shot learning and adaptation: New challenges of support-query shift. In N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, & J. A. Lozano (Eds.), *Machine learning and knowledge discovery in databases. research track - european conference, ECML PKDD 2021, bilbao, spain, september 13-17, 2021, proceedings, part I* (Vol. 12975, pp. 554–569). Springer. Retrieved from https://doi.org/10.1007/978-3-030-86486-6_34 doi: 10.1007/978-3-030-86486-6_34

Bozinovski, S., & Fulgosi, A. (1976). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of symposium informatica* (Vol. 3, pp. 121–126).

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a siamese time delay neural network. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems 6, [7th NIPS conference, denver, colorado, usa, 1993]* (pp. 737–744). Morgan Kaufmann. Retrieved from http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network

Cai, W., Zhang, Y., & Zhou, J. (2013). Maximizing expected model change for active learning in regression. In H. Xiong, G. Karypis, B. Thuraisingham, D. J. Cook, & X. Wu (Eds.), *2013 IEEE 13th international conference on data mining, dallas, tx, usa, december 7-10, 2013* (pp. 51–60). IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICDM.2013.104 doi: 10.1109/ICDM.2013.104

Chitta, K., Alvarez, J. M., & Lesnikowski, A. (2018). Large-scale visual active learning with deep probabilistic ensembles. *CoRR*, *abs/1811.03575*. Retrieved from http://arxiv.org/abs/1811.03575

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. In M. M. Veloso & S. Kambhampati (Eds.), *Proceedings, the twentieth national conference on artificial intelligence and the seventeenth innovative applications of artificial intelligence conference, july 9-13, 2005, pittsburgh, pennsylvania, USA* (pp. 746–751). AAAI Press / The MIT Press. Retrieved from http://www.aaai.org/Library/AAAI/2005/aaai05-117.php

Dagan, I., & Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. In A. Prieditis & S. Russell (Eds.), *Machine learning, proceedings of the twelfth international conference on machine learning, tahoe city, california, usa, july 9-12, 1995* (pp. 150–157). Morgan Kaufmann. Retrieved from https://doi.org/10.1016/b978-1-55860-377-6.50027-x doi: 10.1016/b978-1-55860-377-6.50027-x

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n19-1423 doi: 10.18653/v1/n19-1423

Ding, N., Xu, G., Chen, Y., Wang, X., Han, X., Xie, P., ... Liu, Z. (2021). Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th annual meeting* *of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL/IJCNLP 2021, (volume 1: Long papers), virtual event, august 1-6, 2021* (pp. 3198–3213). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2021.acl-long.248 doi: 10.18653/v1/2021.acl-long.248

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In M. Najork, A. Z. Broder, & S. Chakrabarti (Eds.), *Proceedings of the international conference on web search and web data mining, WSDM 2008, palo alto, california, usa, february 11-12, 2008* (pp. 231–240). ACM. Retrieved from https://doi.org/10.1145/1341531.1341561 doi: 10.1145/1341531.1341561

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., & et al. (2023). A survey for in-context learning. *CoRR*, *abs/2301.00234*. Retrieved from https://doi.org/10.48550/arXiv.2301.00234 doi: 10.48550/arXiv.2301.00234

Dor, L. E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., ... Slonim, N. (2020). Active learning for bert: an empirical study. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 7949–7962).

Ducoffe, M., & Precioso, F. (2018). Adversarial active learning for deep networks: a margin based approach. *CoRR*, *abs/1802.09841*. Retrieved from http://arxiv.org/abs/1802.09841

Edwards, H., & Storkey, A. J. (2016). Towards a neural statistician. *CoRR*, *abs/1606.02185*. Retrieved from http://arxiv.org/abs/1606.02185

Fei-Fei, L., Fergus, R., & Perona, P. (2003). A bayesian approach to unsupervised one-shot learning of object categories. In *9th IEEE international conference on computer vision (ICCV 2003), 14-17 october 2003, nice, france* (pp. 1134–1141). IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICCV.2003.1238476 doi: 10.1109/ICCV.2003.1238476

Fink, M. (2004). Object classification from a single example utilizing class relevance metrics. In *Advances in neural information processing systems 17 [neural information processing* *systems, NIPS 2004, december 13-18, 2004, vancouver, british columbia, canada]* (pp. 449–456). Retrieved from https://proceedings.neurips.cc/paper/2004/hash/ef1e491a766ce3127556063d49bc2f98-Abstract.html

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning, ICML 2017, sydney, nsw, australia, 6-11 august 2017* (Vol. 70, pp. 1126–1135). PMLR. Retrieved from http://proceedings.mlr.press/v70/finn17a.html

Gal, Y., & Ghahramani, Z. (2016a). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd international conference on machine learning, ICML 2016, new york city, ny, usa, june 19-24, 2016* (Vol. 48, pp. 1050–1059). JMLR.org. Retrieved from http://proceedings.mlr.press/v48/gal16.html

Gal, Y., & Ghahramani, Z. (2016b). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd international conference on machine learning, ICML 2016, new york city, ny, usa, june 19-24, 2016* (Vol. 48, pp. 1050–1059). JMLR.org. Retrieved from http://proceedings.mlr.press/v48/gal16.html

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. In *Proceedings of the 34th international conference on machine learning,ICML 2017, sydney, nsw, australia, 6-11 august 2017* (Vol. 70, pp. 1183–1192). PMLR. Retrieved from http://proceedings.mlr.press/v70/gal17a.html

Geifman, Y., & El-Yaniv, R. (2017). Deep active learning over the long tail. *CoRR*, *abs/1711.00941*. Retrieved from http://arxiv.org/abs/1711.00941

Ghorbani, A., Zou, J., & Esteva, A. (2022). Data shapley valuation for efficient batch active learning. In *56th asilomar conference on signals, systems, and computers, ACSSC 2022, pacific grove, ca, usa, october 31 - nov. 2, 2022* (pp. 1456–1462). IEEE. Retrieved from https://doi.org/10.1109/IEEECONF56349.2022.10064696 doi: 10.1109/IEEECONF56349.2022.10064696

Ghorbani, A., & Zou, J. Y. (2019). Data shapley: Equitable valuation of data for machine learning. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference* *on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA* (Vol. 97, pp. 2242–2251). PMLR. Retrieved from http://proceedings.mlr.press/v97/ghorbani19c.html

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2005). Query by committee made real. In *Advances in neural information processing systems 18 [neural information processing systems, NIPS 2005, december 5-8, 2005, vancouver, british columbia, canada]* (pp. 443–450). Retrieved from https://proceedings.neurips.cc/paper/2005/hash/340a39045c40d50dda207bcfdece883a-Abstract.html

Gissin, D., & Shalev-Shwartz, S. (2019). Discriminative active learning. *CoRR*, *abs/1907.06347*. Retrieved from http://arxiv.org/abs/1907.06347

Goldberger, J., Roweis, S. T., Hinton, G. E., & Salakhutdinov, R. (2004). Neighbourhood components analysis. In *Advances in neural information processing systems 17 [neural information processing systems, NIPS 2004, december 13-18, 2004, vancouver, british columbia, canada]* (pp. 513–520). Retrieved from https://proceedings.neurips.cc/paper/2004/hash/42fe880812925e520249e808937738d2-Abstract.html

Goldblum, M., Fowl, L., & Goldstein, T. (2020). Adversarially robust few-shot learning: A meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, neurips 2020, december 6-12, 2020, virtual.* Retrieved from https://proceedings.neurips.cc/paper/2020/hash/cfee398643cbc3dc5eefc89334cacdc1-Abstract.html

Grießhaber, D., Maucher, J., & Vu, N. T. (2020, December). Fine-tuning BERT for low-resource natural language understanding via active learning. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1158–1171). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.100 doi: 10.18653/v1/2020.coling-main.100

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning, ICML 2017, sydney,* *nsw, australia, 6-11 august 2017* (Vol. 70, pp. 1321–1330). PMLR. Retrieved from http://proceedings.mlr.press/v70/guo17a.html

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & et al. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Informatics*, *45*(5), 885–892. Retrieved from https://doi.org/10.1016/j.jbi.2012.04.008 doi: 10.1016/j.jbi.2012.04.008

He, T., Jin, X., Ding, G., Yi, L., & Yan, C. (2019). Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *IEEE international conference on multimedia and expo, ICME 2019, shanghai, china, july 8-12, 2019* (pp. 1360–1365). IEEE. Retrieved from https://doi.org/10.1109/ICME.2019.00236 doi: 10.1109/ICME.2019.00236

Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to learn using gradient descent. In *Artificial neural networks - ICANN 2001, international conference vienna, austria, august 21-25, 2001 proceedings* (Vol. 2130, pp. 87–94). Springer. Retrieved from https://doi.org/10.1007/3-540-44668-0_13 doi: 10.1007/3-540-44668-0_13

Houlsby, N., Huszar, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *CoRR*, *abs/1112.5745*. Retrieved from http://arxiv.org/abs/1112.5745

Huang, J., Child, R., Rao, V., Liu, H., Satheesh, S., & Coates, A. (2016). Active learning for speech recognition: the power of gradients. *CoRR*, *abs/1612.03226*. Retrieved from http://arxiv.org/abs/1612.03226

Jiang, Z., Gao, Z., Duan, Y., Kang, Y., Sun, C., & et al. (2020). Camouflaged chinese spam content detection with semi-supervised generative active learning. In *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, online, july 5-10, 2020* (pp. 3080–3085). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.acl-main.279 doi: 10.18653/v1/2020.acl-main.279

Kim, K., Park, D., Kim, K. I., & Chun, S. Y. (2021). Task-aware variational adversarial active learning. In *IEEE conference on computer vision and pattern recognition, CVPR 2021, virtual, june 19-25, 2021* (pp. 8166–8175). Computer Vision Foundation / IEEE. Retrieved from https://openaccess.thecvf.com/content/CVPR2021/html/Kim_Task-Aware_Variational_Adversarial_Active_Learning_CVPR_2021_paper.html doi: 10.1109/CVPR46437.2021.00807

Kirsch, A., van Amersfoort, J., & Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 7024–7035). Retrieved from https://proceedings.neurips.cc/paper/2019/hash/95323660ed2124450caaac2c46b5ed90-Abstract.html

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *Icml deep learning workshop* (Vol. 2).

Köksal, A., Schick, T., & Schütze, H. (2022). MEAL: stable and active learning for few-shot prompting. *CoRR*, *abs/2211.08358*. Retrieved from https://doi.org/10.48550/arXiv.2211.08358 doi: 10.48550/arXiv.2211.08358

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. proceedings of a meeting held december 3-6, 2012, lake tahoe, nevada, united states* (pp. 1106–1114). Retrieved from https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kwok, K., Grunfeld, L., Dinstl, N., & Chan, M. (2000). TREC-9 cross language, web and question-answering track experiments using PIRCS. In *Proceedings of the ninth text retrieval conference, TREC 2000, gaithersburg, maryland, usa, november 13-16, 2000* (Vol. 500-249). National Institute of Standards and Technology (NIST). Retrieved from http://trec.nist.gov/pubs/trec9/papers/pircst9.pdf

Lewis, D. D., & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning.

In *Machine learning, proceedings of the eleventh international conference, rutgers university, new brunswick, nj, usa, july 10-13, 1994* (pp. 148–156). Morgan Kaufmann. Retrieved from https://doi.org/10.1016/b978-1-55860-335-6.50026-x doi: 10.1016/b978-1-55860-335-6.50026-x

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM-SIGIR conference on research and development in information retrieval. dublin, ireland, 3-6 july 1994 (special issue of the SIGIR forum)* (pp. 3–12). ACM/Springer. Retrieved from https://doi.org/10.1007/978-1-4471-2099-5_1 doi: 10.1007/978-1-4471-2099-5_1

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, H., Dong, W., Mei, X., Ma, C., Huang, F., & Hu, B. (2019). Lgm-net: Learning to generate matching networks for few-shot learning. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA* (Vol. 97, pp. 3825–3834). PMLR. Retrieved from http://proceedings.mlr.press/v97/li19c.html

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., & et al. title = Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning, v. . a. y. . . u. . h. d. . a., journal = CoRR. (n.d.).

Liu, P., Zhang, H., & Eom, K. B. (2017). Active deep learning for classification of hyperspectral images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, *10*(2), 712–724. Retrieved from https://doi.org/10.1109/JSTARS.2016.2598859 doi: 10.1109/JSTARS.2016.2598859

Lu, J., Gong, P., Ye, J., & Zhang, C. (2020). Learning from very few samples: A survey. *CoRR*, *abs/2009.02653*. Retrieved from https://arxiv.org/abs/2009.02653

Mahabadi, R. K., Zettlemoyer, L., Henderson, J., Saeidi, M., Mathias, L., Stoyanov, V., & Yazdani,

M. (2022). PERFECT: prompt-free and efficient few-shot learning with language models. *CoRR*, *abs/2204.01172*. Retrieved from https://doi.org/10.48550/arXiv.2204.01172 doi: 10.48550/arXiv.2204.01172

Marbinah, J. (2021). *Hybrid pool based deep active learning for object detection using intermediate network embeddings.*

Margatina, K., Vernikos, G., Barrault, L., & Aletras, N. (2021, November). Active learning by acquiring contrastive examples. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 650–663). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.51 doi: 10.18653/v1/2021.emnlp-main.51

McCallum, A., & Nigam, K. (1998). Employing EM and pool-based active learning for text classification. In J. W. Shavlik (Ed.), *Proceedings of the fifteenth international conference on machine learning (ICML 1998), madison, wisconsin, usa, july 24-27, 1998* (pp. 350–358). Morgan Kaufmann.

Mensink, T., Verbeek, J., Perronnin, F., & Csurka, G. (2013). Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(11), 2624–2637. Retrieved from https://doi.org/10.1109/TPAMI.2013.83 doi: 10.1109/TPAMI.2013.83

Messaoudi, C., Guessoum, Z., & Ben Romdhane, L. (2022). Opinion mining in online social media: a survey. *Social Network Analysis and Mining*, *12*(1), 25.

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes - which naive bayes? In *CEAS 2006 - the third conference on email and anti-spam, july 27-28, 2006, mountain view, california, USA*. Retrieved from http://www.ceas.cc/2006/listabs.html#15.pdf

Miller, E. G., Matsakis, N. E., & Viola, P. A. (2000). Learning from one example through shared densities on transforms. In *2000 conference on computer vision and pattern recognition (CVPR 2000), 13-15 june 2000, hilton head, sc, USA* (pp. 1464–1471). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2000.855856 doi: 10.1109/CVPR.2000.855856

Mottaghi, A., & Yeung, S. (2019). Adversarial representation active learning. *CoRR*, *abs/1912.09720*. Retrieved from http://arxiv.org/abs/1912.09720

Müller, T., Pérez-Torró, G., Basile, A., & Franco-Salvador, M. (2022). Active few-shot learning with FASL. In *Natural language processing and information systems - 27th international conference on applications of natural language to information systems, NLDB 2022, valencia, spain, june 15-17, 2022, proceedings* (Vol. 13286, pp. 98–110). Springer. Retrieved from https://doi.org/10.1007/978-3-031-08473-7_9 doi: 10.1007/978-3-031-08473-7_9

Müller, T., Pérez-Torró, G., & Franco-Salvador, M. (2022). Few-shot learning with siamese networks and label tuning. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2022, dublin, ireland, may 22-27, 2022* (pp. 8532–8545). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2022.acl-long.584 doi: 10.18653/v1/2022.acl-long.584

Musgrave, K., Belongie, S. J., & Lim, S. (2020). A metric learning reality check. In *Computer vision - ECCV 2020 - 16th european conference, glasgow, uk, august 23-28, 2020, proceedings, part XXV* (Vol. 12370, pp. 681–699). Springer. Retrieved from https://doi.org/10.1007/978-3-030-58595-2_41 doi: 10.1007/978-3-030-58595-2_41

Mussmann, S., Reisler, J., Tsai, D., Mousavi, E., O'Brien, S., & et al. (2022). Active learning with expected error reduction. *CoRR*, *abs/2211.09283*. Retrieved from https://doi.org/10.48550/arXiv.2211.09283 doi: 10.48550/arXiv.2211.09283

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). *doccano: Text annotation tool for human.* Retrieved from https://github.com/doccano/doccano (Software available from https://github.com/doccano/doccano)

Neves, M. L., & Leser, U. (2014). A survey on annotation tools for the biomedical literature. *Briefings Bioinform.*, *15*(2), 327–340. Retrieved from https://doi.org/10.1093/bib/bbs084 doi: 10.1093/bib/bbs084

Omar, R., Dhall, I., Kalnis, P., & Mansour, E. (2023). A universal question-answering platform for knowledge graphs. *Proc. ACM Manag. Data*, *1*(1), 57:1–57:25. Retrieved from https://

doi.org/10.1145/3588911 doi: 10.1145/3588911

Ostapuk, N., Yang, J., & Cudré-Mauroux, P. (2019). Activelink: Deep active learning for link prediction in knowledge graphs. In *The world wide web conference, WWW 2019, san francisco, ca, usa, may 13-17, 2019* (pp. 1398–1408). ACM. Retrieved from https://doi.org/10.1145/3308558.3313620 doi: 10.1145/3308558.3313620

Pei, J., Ananthasubramaniam, A., Wang, X., Zhou, N., Sargent, J., Dedeloudis, A., & Jurgens, D. (2022). Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*.

Perry, T. (2021, November). LightTag: Text annotation platform. In *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations* (pp. 20–27). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-demo.3 doi: 10.18653/v1/2021.emnlp-demo.3

Pezeshkpour, P., Zhao, Z., & Singh, S. (2020). On the utility of active instance selection for few-shot learning. *NeurIPS HAMLETS*.

Pop, R., & Fulop, P. (2018). Deep ensemble bayesian active learning : Addressing the mode collapse issue in monte carlo dropout via ensembles. *CoRR*, *abs/1811.03897*. Retrieved from http://arxiv.org/abs/1811.03897

Prodigy. (2017). Retrieved from https://prodi.gy

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Qiu, Z., Miller, D. J., & Kesidis, G. (2017). A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes. *IEEE Trans. Neural Networks Learn. Syst.*, *28*(4), 917–933. Retrieved from https://doi.org/10.1109/TNNLS.2016.2514401 doi: 10.1109/TNNLS.2016.2514401

Radmard, P., Fathullah, Y., & Lipani, A. (2021). Subsequence based deep active learning for named entity recognition. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL/IJCNLP 2021, (volume 1: Long* *papers), virtual event, august 1-6, 2021* (pp. 4310–4321). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2021.acl-long.332 doi: 10.18653/v1/2021.acl-long.332

Rahaman, R., & Thiéry, A. H. (2021). Uncertainty quantification and deep ensembles. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, neurips 2021, december 6-14, 2021, virtual* (pp. 20063–20075). Retrieved from https://proceedings.neurips.cc/paper/2021/hash/a70dc40477bc2adceef4d2c90f47eb82-Abstract.html

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., & et al. (2021). Speechbrain: A general-purpose speech toolkit. *CoRR*, *abs/2106.04624*. Retrieved from https://arxiv.org/abs/2106.04624

Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.* OpenReview.net. Retrieved from https://openreview.net/forum?id=rJY0-Kcll

Rebuffi, S., Ehrhardt, S., Han, K., Vedaldi, A., & Zisserman, A. (2020). Semi-supervised learning with scarce annotations. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR workshops 2020, seattle, wa, usa, june 14-19, 2020* (pp. 3294–3302). Computer Vision Foundation / IEEE. Retrieved from https://openaccess.thecvf.com/content_CVPRW_2020/html/w45/Rebuffi_Semi-Supervised_Learning_With_Scarce_Annotations_CVPRW_2020_paper.html doi: 10.1109/CVPRW50498.2020.00389

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, hong kong, china, november 3-7, 2019* (pp. 3980–3990). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/D19-1410 doi: 10.18653/v1/D19-1410

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., … Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, *54*(9), 1–40.

Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K., & Wierstra, D. (2016). One-shot generalization in deep generative models. In *Proceedings of the 33nd international conference on machine learning, ICML 2016, new york city, ny, usa, june 19-24, 2016* (Vol. 48, pp. 1521–1529). JMLR.org. Retrieved from http://proceedings.mlr.press/v48/rezende16.html

Roy, N., & McCallum, A. (2001a). Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, *2*, 441–448.

Roy, N., & McCallum, A. (2001b). Toward optimal active learning through sampling estimation of error reduction. In C. E. Brodley & A. P. Danyluk (Eds.), *Proceedings of the eighteenth international conference on machine learning (ICML 2001), williams college, williamstown, ma, usa, june 28 - july 1, 2001* (pp. 441–448). Morgan Kaufmann.

Salakhutdinov, R., & Hinton, G. E. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of the eleventh international conference on artificial intelligence and statistics, AISTATS 2007, san juan, puerto rico, march 21-24, 2007* (Vol. 2, pp. 412–419). JMLR.org. Retrieved from http://proceedings.mlr.press/v2/salakhutdinov07a.html

Sang, E. F. T. K., & Meulder, F. D. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on natural language learning, conll 2003, held in cooperation with HLT-NAACL 2003, edmonton, canada, may 31 - june 1, 2003* (pp. 142–147). ACL. Retrieved from https://aclanthology.org/W03-0419/

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., & et al. (2022). Multitask prompted training enables zero-shot task generalization. In *The tenth international conference on learning representations, ICLR 2022, virtual event, april 25-29, 2022.* OpenReview.net. Retrieved from https://openreview.net/forum?id=9Vrb9D0WI4

Saquil, Y., Kim, K. I., & Hall, P. M. (2018). Ranking cgans: Subjective control over semantic image attributes. In *British machine vision conference 2018, BMVC 2018, newcastle, uk, september*

3-6, 2018 (p. 131). BMVA Press. Retrieved from http://bmvc2018.org/contents/papers/0534.pdf

Satorras, V. G., & Estrach, J. B. (2018). Few-shot learning with graph neural networks. In *6th international conference on learning representations, ICLR 2018, vancouver, bc, canada, april 30 - may 3, 2018, conference track proceedings.* OpenReview.net. Retrieved from https://openreview.net/forum?id=BJj6qGbRW

Scheffer, T., Decomain, C., & Wrobel, S. (2001). Active hidden markov models for information extraction. In *Advances in intelligent data analysis, 4th international conference, IDA 2001, cascais, portugal, september 13-15, 2001, proceedings* (Vol. 2189, pp. 309–318). Springer. Retrieved from https://doi.org/10.1007/3-540-44816-0_31 doi: 10.1007/3-540-44816-0_31

Schein, A. I., & Ungar, L. H. (2007). Active learning for logistic regression: an evaluation. *Mach. Learn.*, *68*(3), 235–265. Retrieved from https://doi.org/10.1007/s10994-007-5019-5 doi: 10.1007/s10994-007-5019-5

Schröder, C., Müller, L., Niekler, A., & Potthast, M. (2023). Small-text: Active learning for text classification in python. In D. Croce & L. Soldaini (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics. EACL 2023 - system demonstrations, dubrovnik, croatia, may 2-4, 2023* (pp. 84–95). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.eacl-demo.11

Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *6th international conference on learning representations, ICLR 2018, vancouver, bc, canada, april 30 - may 3, 2018, conference track proceedings.* OpenReview.net. Retrieved from https://openreview.net/forum?id=H1aIuk-RW

Settles, B. (2009). Active learning literature survey.

Settles, B. (2011). From theories to queries. In *Active learning and experimental design workshop, in conjunction with AISTATS 2010, sardinia, italy, may 16, 2010* (Vol. 16, pp. 1–18). JMLR.org. Retrieved from http://proceedings.mlr.press/v16/settles11a/settles11a.pdf

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *2008 conference on empirical methods in natural language processing, EMNLP 2008, proceedings of the conference, 25-27 october 2008, honolulu, hawaii, usa, A meeting of sigdat, a special interest group of the ACL* (pp. 1070–1079). ACL. Retrieved from https://aclanthology.org/D08-1112/

Settles, B., Craven, M., & Ray, S. (2007). Multiple-instance active learning. *Advances in neural information processing systems*, *20*.

Seung, H. S., Opper, M., & Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth annual ACM conference on computational learning theory, COLT 1992, pittsburgh, pa, usa, july 27-29, 1992* (pp. 287–294). ACM. Retrieved from https://doi.org/10.1145/130385.130417 doi: 10.1145/130385.130417

Shelmanov, A., Puzyrev, D., Kupriyanova, L., Belyakov, D., Larionov, D., & et al. (2021). Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume, EACL 2021, online, april 19 - 23, 2021* (pp. 1698–1712). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2021.eacl-main.145 doi: 10.18653/v1/2021.eacl-main.145

Shen, Y., Yun, H., Lipton, Z. C., Kronrod, Y., & Anandkumar, A. (2017). Deep active learning for named entity recognition. In *Proceedings of the 2nd workshop on representation learning for nlp, rep4nlp@acl 2017, vancouver, canada, august 3, 2017* (pp. 252–256). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/w17-2630 doi: 10.18653/v1/w17-2630

Shnarch, E., Halfon, A., Gera, A., Danilevsky, M., Katsis, Y., Choshen, L., … others (2022). Label sleuth: From unlabeled text to a classifier in a few hours. *arXiv preprint arXiv:2208.01483*.

Shu, J., Xu, Z., & Meng, D. (2018). Small sample learning in big data era. *CoRR*, *abs/1808.04572*. Retrieved from http://arxiv.org/abs/1808.04572

Shui, C., Zhou, F., Gagné, C., & Wang, B. (2020). Deep active learning: Unified and principled method for query and training. In S. Chiappa & R. Calandra (Eds.), *The 23rd international conference on artificial intelligence and statistics, AISTATS 2020, 26-28 august*

2020, online [palermo, sicily, italy] (Vol. 108, pp. 1308–1318). PMLR. Retrieved from http://proceedings.mlr.press/v108/shui20a.html

Siddhant, A., & Lipton, Z. C. (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, brussels, belgium, october 31 - november 4, 2018* (pp. 2904–2909). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/d18-1318 doi: 10.18653/v1/d18-1318

Siméoni, O., Budnik, M., Avrithis, Y., & Gravier, G. (2020). Rethinking deep active learning: Using unlabeled data at model training. In *25th international conference on pattern recognition, ICPR 2020, virtual event / milan, italy, january 10-15, 2021* (pp. 1220–1227). IEEE. Retrieved from https://doi.org/10.1109/ICPR48806.2021.9412716 doi: 10.1109/ICPR48806.2021.9412716

Sinha, S., Ebrahimi, S., & Darrell, T. (2019). Variational adversarial active learning. In *2019 IEEE/CVF international conference on computer vision, ICCV 2019, seoul, korea (south), october 27 - november 2, 2019* (pp. 5971–5980). IEEE. Retrieved from https://doi.org/10.1109/ICCV.2019.00607 doi: 10.1109/ICCV.2019.00607

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 4077–4087). Retrieved from https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html

Snow, R., O'connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 254–263).

Song, Y., Wang, T., Mondal, S. K., & Sahoo, J. P. (2022). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *CoRR*, *abs/2205.06743*. Retrieved from https://doi.org/10.48550/arXiv.2205.06743 doi: 10.48550/arXiv.2205.06743

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (pp. 3104–3112). Retrieved from https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Tam, D., Menon, R. R., Bansal, M., Srivastava, S., & Raffel, C. (2021). Improving and simplifying pattern exploiting training. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing, EMNLP 2021, virtual event / punta cana, dominican republic, 7-11 november, 2021* (pp. 4980–4991). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2021.emnlp-main.407 doi: 10.18653/v1/2021.emnlp-main.407

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020-2022). *Label Studio: Data labeling software.* Retrieved from https://github.com/heartexlabs/label-studio (Open source software available from https://github.com/heartexlabs/label-studio)

Trivedi, P., Maheshwari, G., Dubey, M., & Lehmann, J. (2017). Lc-quad: A corpus for complex question answering over knowledge graphs. In *The semantic web - ISWC 2017 - 16th international semantic web conference, vienna, austria, october 21-25, 2017, proceedings, part II* (Vol. 10588, pp. 210–218). Springer. Retrieved from https://doi.org/10.1007/978-3-319-68204-4_22 doi: 10.1007/978-3-319-68204-4_22

Tseng, H., Lee, H., Huang, J., & Yang, M. (2020). Cross-domain few-shot classification via learned feature-wise transformation. In *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, april 26-30, 2020.* OpenReview.net. Retrieved from https://openreview.net/forum?id=SJl5Np4tPr

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *CoRR*, *abs/2209.11055*. Retrieved from https://doi.org/10.48550/arXiv.2209.11055 doi: 10.48550/arXiv.2209.11055

Tür, G., Hakkani-Tür, D., & Schapire, R. E. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Commun.*, *45*(2), 171–186. Retrieved from https://doi.org/10.1016/j.specom.2004.08.002 doi: 10.1016/j.specom.2004.08.002

Usbeck, R., Gusmita, R. H., Ngomo, A. N., & Saleem, M. (2018). 9th challenge on question answering over linked data (QALD-9) (invited paper). In *Joint proceedings of the 4th workshop on semantic deep learning (semdeep-4) and nliwod4: Natural language interfaces for the web of data (NLIWOD-4) and 9th question answering over linked data challenge (QALD-9) co-located with 17th international semantic web conference (ISWC 2018), monterey, california, united states of america, october 8th - 9th, 2018* (Vol. 2241, pp. 58–64). CEUR-WS.org. Retrieved from https://ceur-ws.org/Vol-2241/paper-06.pdf

Varma, P., & Ré, C. (2018). Snuba: Automating weak supervision to label training data. In *Proceedings of the vldb endowment. international conference on very large data bases* (Vol. 12, p. 223).

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016, december 5-10, 2016, barcelona, spain* (pp. 3630–3638). Retrieved from https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

Vlachos, A. (2008). A stopping criterion for active learning. *Comput. Speech Lang.*, *22*(3), 295–312. Retrieved from https://doi.org/10.1016/j.csl.2007.12.001 doi: 10.1016/j.csl.2007.12.001

Wang, T., Zhao, X., Lv, Q., Hu, B., & Sun, D. (2021). Density weighted diversity based query strategy for active learning. In *24th IEEE international conference on computer supported cooperative work in design, CSCWD 2021, dalian, china, may 5-7, 2021* (pp. 156–161). IEEE. Retrieved from https://doi.org/10.1109/CSCWD49262.2021.9437695 doi: 10.1109/CSCWD49262.2021.9437695

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, *53*(3), 63:1–63:34. Retrieved from https://doi.org/10.1145/3386252 doi: 10.1145/3386252

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., & et al. (2022). Chain of thought prompting elicits reasoning in large language models. *CoRR*, *abs/2201.11903*. Retrieved from https://arxiv.org/abs/2201.11903

Wolf, L., & Martin, I. (2005a). Robust boosting for learning from few examples. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005), 20-26 june 2005, san diego, ca, USA* (pp. 359–364). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2005.305 doi: 10.1109/CVPR.2005.305

Wolf, L., & Martin, I. (2005b). Robust boosting for learning from few examples. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR 2005), 20-26 june 2005, san diego, ca, USA* (pp. 359–364). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2005.305 doi: 10.1109/CVPR.2005.305

Woodward, M., & Finn, C. (2017). Active one-shot learning. *CoRR*, *abs/1702.06559*. Retrieved from http://arxiv.org/abs/1702.06559

Yang, Y., & Katiyar, A. (2020a). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 6365–6375). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-main.516 doi: 10.18653/v1/2020.emnlp-main.516

Yang, Y., & Katiyar, A. (2020b, November). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 6365–6375). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.516 doi: 10.18653/v1/2020.emnlp-main.516

Yin, C., Qian, B., Cao, S., Li, X., Wei, J., & et al. (2017). Deep similarity-based batch mode active learning with exploration-exploitation. In *2017 IEEE international conference on data mining, ICDM 2017, new orleans, la, usa, november 18-21, 2017* (pp. 575–584). IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICDM.2017.67 doi: 10.1109/ICDM.2017.67

Yin, C., Qian, B., Cao, S., Li, X., Wei, J., Zheng, Q., & Davidson, I. (2017). Deep similarity-based batch mode active learning with exploration-exploitation. In *2017 IEEE international conference on data mining, ICDM 2017, new orleans, la, usa, november 18-21, 2017* (pp. 575–584). IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICDM.2017.67 doi: 10.1109/ICDM.2017.67

Yu, M., Guo, X., Yi, J., Chang, S., Potdar, S., & et al. (2018). Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2018, new orleans, louisiana, usa, june 1-6, 2018, volume 1 (long papers)* (pp. 1206–1215). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/n18-1109 doi: 10.18653/v1/n18-1109

Yuan, M., Lin, H., & Boyd-Graber, J. L. (2020). Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, online, november 16-20, 2020* (pp. 7935–7948). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-main.637 doi: 10.18653/v1/2020.emnlp-main.637

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1476–1485).

Zhang, C., & Chen, T. (2002). An active learning framework for content-based information retrieval. *IEEE Trans. Multim.*, *4*(2), 260–268. Retrieved from https://doi.org/10.1109/TMM.2002.1017738 doi: 10.1109/TMM.2002.1017738

Zhang, T., & Oles, F. (2000). *A probability analysis on the value of unlabeled data for classification problems. 17th icml (pp. 1191–1198).* Stanford, US.

Zhang, Y., Lease, M., & Wallace, B. (2017). Active discriminative text representation learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Zhdanov, F. (2019a). Diverse mini-batch active learning. *CoRR*, *abs/1901.05954*. Retrieved from http://arxiv.org/abs/1901.05954

Zhdanov, F. (2019b). Diverse mini-batch active learning. *CoRR*, *abs/1901.05954*. Retrieved from http://arxiv.org/abs/1901.05954

Zhu, J., & Bento, J. (2017). Generative adversarial active learning. *CoRR*, *abs/1702.07956*. Retrieved from http://arxiv.org/abs/1702.07956

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *Icml 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining* (Vol. 3).

Zhu, Z. L., Yadav, V., Afzal, Z., & Tsatsaronis, G. (2022). Few-shot initializing of active learner via meta-learning. In *Findings of the association for computational linguistics: EMNLP 2022, abu dhabi, united arab emirates, december 7-11, 2022* (pp. 1117–1133). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-emnlp.80

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., & et al. (2021). A comprehensive survey on transfer learning. *Proc. IEEE*, *109*(1), 43–76. Retrieved from https://doi.org/10.1109/JPROC.2020.3004555 doi: 10.1109/JPROC.2020.3004555