

ABSTRACT

As email has become a primary channel for customer service and operational workflows, organizations face substantial costs in handling large volumes of incoming messages. Traditional model-based classification is widely recognized as an effective approach for organizing such emails. However, these systems typically require a large, manually labeled training set, which creates a high upfront cost. To address this challenge, this thesis proposes an active learning-based pipeline for sequential class discovery and analyzes how effectively it can uncover the class structure present in the training data while improving classification performance under limited labeling budgets.

To achieve this goal, the proposed pipeline is evaluated through combinations of key components. We compare multiple query strategies designed to prioritize informative samples, and we investigate clustering-based seed sampling strategies intended to improve early-stage learning. As base classifiers, we apply active learning to traditional models such as k-nearest neighbors and logistic regression. We also include a Retrieval Augmented Generation (RAG)-based classifier that can achieve strong predictive performance. The evaluation focuses on class discovery dynamics throughout training, including the speed of discovering the full set of classes and the time required to identify minority classes, as well as the changes in classification performance across labeling iterations.

Overall, the experimental results show that active learning substantially improves label efficiency compared to random sampling. Across settings, the pipeline discovers the overall class structure earlier with fewer labeled instances, identifies minority classes sooner, and reaches target performance levels using fewer labels. In contrast, clustering-based seed sampling yields only small and inconsistent gains over random seeding, making it difficult to justify its added complexity of the pipeline. When combined with active learning, the RAG-based classifier discovers classes faster than random sampling and achieves slightly higher classification performance, and it also outperforms traditional classifiers used within the same active learning framework. However, the performance gains are modest relative to the implementation effort and operational cost of integrating RAG with active learning, suggesting that its practical value in real-world deployments requires further discussion.

These findings highlight that an active learning pipeline can explore class structure effectively at the early labeling stage and achieve higher classification performance than random labeling under the same labeling budget.

Keywords: Active Learning, Sequential Class Discovery, Email Intent Classification, Label Efficiency, Query Strategies

ZUSAMMENFASSUNG

Die zunehmende Nutzung von E-Mail im Kundenservice und in operativen Workflows verursacht in vielen Organisationen erhebliche Kosten bei der Verarbeitung großer Mengen eingehender Nachrichten. Modellbasierte Klassifikation kann diese E-Mails zwar effektiv sortieren, erfordert jedoch häufig umfangreiche manuelle Annotationen und damit hohe Anfangskosten. Diese Arbeit schlägt eine Active-Learning-basierte Pipeline zur sequenziellen Klassenentdeckung vor und untersucht, wie gut sich unter begrenztem Labeling-Budget die Klassenstruktur erschließen lässt und wie sich die Klassifikationsleistung im Trainingsverlauf entwickelt.

Zur Erreichung dieses Ziels wird die Pipeline in unterschiedlichen Kombinationen ihrer zentralen Komponenten evaluiert. Hierzu werden Query-Strategien zur Auswahl informativer Beispiele und clusterbasierte Seed-Sampling-Strategien für die frühe Lernphase verglichen. Als Basis-Klassifikatoren werden k-nächste Nachbarn und logistische Regression eingesetzt. Zusätzlich berücksichtigen wir einen RAG-basierten Klassifikator, der eine hohe Vorhersagequalität erzielen kann. Die Evaluation fokussiert die Dynamik der Klassenentdeckung im Trainingsverlauf, darunter die Geschwindigkeit, mit der der vollständige Klassensatz entdeckt wird, sowie die benötigte Zeit zur Identifikation von Minderheitsklassen. Ergänzend wird die Entwicklung der Klassifikationsleistung im Trainingsverlauf analysiert.

Die Ergebnisse zeigen, dass Active Learning die Label-Effizienz gegenüber zufälligem Sampling deutlich erhöht. Über verschiedene Einstellungen hinweg entdeckt die Pipeline die globale Klassenstruktur früher mit weniger gelabelten Instanzen, identifiziert Minderheitsklassen schneller und erreicht Zielwerte der Performanz mit weniger Annotationen. Clusterbasiertes Seed-Sampling liefert dagegen nur geringe und inkonsistente Zugewinne und rechtfertigt die zusätzliche Komplexität kaum. In Kombination mit Active Learning entdeckt der RAG-basierte Klassifikator Klassen schneller als bei random Sampling und erzielt eine leicht höhere Klassifikationsleistung. Zudem übertrifft die traditionellen Modelle im selben Rahmen. Insgesamt bleiben die Zugewinne jedoch gemessen am Implementierungsaufwand und den operativen Kosten einer RAG-Integration moderat, sodass der praktische Mehrwert für reale Deployments weiter zu diskutieren ist.

Diese Ergebnisse unterstreichen, dass eine Active-Learning-Pipeline die Klassenstruktur bereits in einer frühen Labeling-Phase effektiv explorieren kann und unter demselben Labeling-Budget eine höhere Klassifikationsleistung als zufälliges Labeling erreicht.