# Active learning for text classification using rationales

Teja Kanchinadam, Rick Lentz, Qian You, Glenn Fung

***Abstract*— Active learning is an iterative process that trains machine learning algorithms while minimizing the need for labeled data. It has two notable components a) a machine learning algorithm which can pick the best samples to train itself b) a human/oracle who can label these samples. The entire goal of the active learning process is to minimize the efforts of the latter part while generalizing well on the unseen data. Our work describes an approach for text classification using Active learning. In this approach we not only pick the best samples to train with but also enrich the document feature space with the rationale within the document. For example a rationale of movie review can be the keywords and phrases that show positive sentiment. Preliminary results indicate that our methodology leads to faster convergence rate than other active learning based classifications.**

## I. INTRODUCTION

Active learning is a subset of machine learning that seeks to optimize the learning efficiency by choosing the best data samples for training. The typical setup for pool-based active learning for classification is: an unlabeled pool of examples $\mathcal{U}$, a labeled pool $\mathcal{L}$ of example-label pairs $(x, y_x)$, an oracle that can supply the label of any $x \in \mathcal{U}$, and a querying strategy that selects which example in $\mathcal{U}$ the oracle should label based on the current state of $\mathcal{L}$. On the other hand, *Passive learning* would just sample uniformly from $\mathcal{U}$ as its querying strategy. In our case, human annotators provide the role of the oracle. Active learning is particularly useful when human annotation is involved, because one can substantially reduce the cost and time needed.

The goal of the querying strategy is to select $x^* \in \mathcal{U}$ such that $\mathcal{L}^* = \mathcal{L} \cup \{(x^*, y_{x^*})\}$ yields the maximum information gain versus $\mathcal{L} \cup \{(x, y_x)\}$ for any other $x \in \mathcal{U}$. In this work, we implement *uncertainty sampling*, a querying strategy that $\mu(x)$ captures the classifier's uncertainty about the class of $x$. The intuition is that not much information is gained if a classifier gets a new label with which it already agreed with high certainty.

There is prior work that uses rationales to improve the active learning process, however, as far as we know, most of the prior work use word frequencies (bag-of-words) high dimensional sparse representations and focus on leveraging the rationales to weight feature importance at the moment of training. In contrast, we are focusing in taking advantage of word-embedding-based representations where not only syntactic but also semantic information is represented in a compact lower dimensional space where active learning algorithms are known to learn faster. Furthermore, words, sentences, and whole documents can be represented in a common space where the semantic relations among them are preserved.

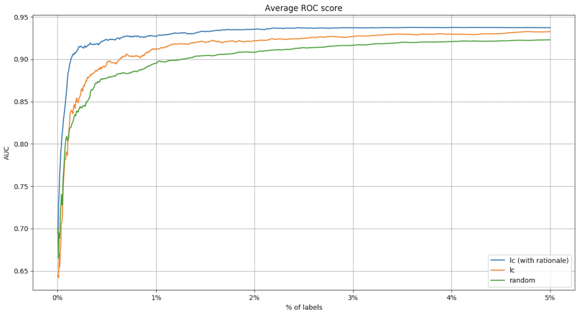

Fig. 1. Active learning simulation results on IMDB movies dataset. In this figure, the least certain (lc) query strategy using the document features with rationale is converging faster, compared to the least certain and random query strategies with only document features.

## II. RATIONALES

In this work, we ask the human annotators to not only label the sample but also to highlight the *rationale* behind their decisions. We then use these rationales as part of our training process to help accelerate the process. We use the rationale information as follows:

First, the document level feature $x_i$ for the document $d_i$ are given as

$$x_i = \varphi_\vartheta\left(d_i\right) \tag{1}$$

where $\varphi_\vartheta$ is Universal Sentence Encoder (USE), an off-the-shelf pretrained model.

Second, we enrich these document level features using rationale $d_r$ for the document $d_i$ as follows:

$$x_r = \varphi_\vartheta\left(d_r\right) \tag{2}$$

$$x_f = \lambda * x_i + (1 - \lambda) * x_r \tag{3}$$

where $\lambda$ is a value in the range $[0, 1]$ and $x_f$ are the enriched features used for training. During inference and query strategy, we only use $x_i$ since we don't have $d_r$.

## III. PRELIMINARY RESULTS

We used the IMDB movie review dataset and gathered rationales from Mechanical Turk by asking humans to highlight the reasons why the review is positive or negative. We then simulate the active learning process with Logistic Regression, a probabilistic model, as our base classifier and the results are shown in Figure 1.

From Figure 1, it is evident that document features enriched with rationale help the active learning process through faster convergence. We are excited with these results and are focusing on making the process more efficient.