Practical Dataless Text Classification Through Dense Retrieval

Anonymous ACL submission

Abstract

Dataless text classification aims to classify documents using only class descriptions without any training data. Recent research shows that pre-trained textual entailment models can achieve state-of-the-art dataless classification performance on various tasks. However, such models are not practical in that their prediction speed is slow as they need k forward passes to predict k classes and they are not built for fine-tuning to further improve the initial (often mediocre) performance. This work proposes a simple, effective, and practical dataless classification approach. We use class descriptions as queries to retrieve task-specific or external unlabeled data on which pseudo-labels are assigned to train a classifier. Experiments on a wide range of classification tasks show that the proposed approach consistently outperforms entailment-based models in terms of classification accuracy, prediction speed, and performance gain when fine-tuned on labeled data.

1 Introduction

Text classification is one of the most used techniques in mining large-scale unstructured text. When sufficient labeled data are available, supervised classification techniques can achieve excellent performance. However, manually labeling example documents can be time-consuming and labor-intensive, a major burden when applying supervised text classification techniques in practice.

Recently, *dataless text classification* (Chang et al., 2008; Druck et al., 2008; Song and Roth, 2014; Chen et al., 2015; Li et al., 2016a,b; Song et al., 2016) has been proposed to save labeling efforts. It refers to the ability for a machine learning model to start classifying documents by using only class descriptions and *no* training data. Since any text classification task necessarily starts with a description for each class, class descriptions are naturally available from the very beginning. Therefore, dataless text classification has practical value in real-world applications.

Early research showed that dataless classifiers are able to classify documents on unbounded label sets if label descriptions are carefully written, e.g., paraphrasing the same concept using different synonyms and from multiple aspects (Chang et al., 2008; Wang and Domeniconi, 2009; Song and Roth, 2014). These approaches often leverage external resources such as Wikipedia to construct semantic representations for both class descriptions and text documents. Many different settings have been considered in previous works, some using slightly different names, including *zero-shot text classification* (Pushp and Srivastava, 2017; Yin et al., 2019) and *weakly supervised text classification* (Chu et al., 2020a). Recent research found that Transformer-based textual entailment models can provide more competitive performance on dataless classification tasks (Yin et al., 2019; Chu et al., 2020a). The basic idea is to ask a pre-trained textual entailment model to judge if a document logically entails any of the class descriptions, and then pick the class with the highest probability of entailment. Such an approach is shown to give better performance than earlier approaches thanks to the contextual text representations learned by deep Transformers such as BERT (Devlin et al., 2018).

However, the textual entailment approach to dataless classification has several drawbacks which diminish its practical value. First, one has to run the entailment model k times to classify one document into k categories. The prediction speed slows down as more categories (larger k) are considered in a task. Second, the performance of a dataless text classifier is often far from optimal, and therefore practitioners often wish to further improve it using labeled examples afterwards. As we will show in the experiments, entailment models are not ideal for fine-tuning on classification tasks. They aim to solve a much harder problem than classification – to learn semantic dependencies between all words in the document and all words in the class definition – and therefore need more data to learn well. Third, it is difficult to adapt a well-trained entailment model on a task-specific corpus, since adaptive pretraining (such as masked language modeling) has to happen *before* the entailment model is trained. Lastly, the performance of entailment-based classifiers tends to vary significantly across different tasks (Yin et al., 2019). Recent work showed that they sometimes even underperform a raw BERT model that is not fine-tuned on entailment tasks (Ma et al., 2021).

Ideally, a dataless text classifier should not only provide a decent performance to jump-start the task, but also be readily adaptable to task-specific unlabeled data, continuously trainable if labeled data ever become available, and scalable to a large number of categories at prediction time. In this paper, we propose methods that achieve these goals. The main idea is to create pseudo-labeled documents for each class using class descriptions as queries and dense retrieval models as pseudo-labeling functions. These pseudo-labeled data are then used to train a classifier. This simple idea has its root in early information retrieval research, such as pseudo-relevance feedback (Rocchio, 1965) and naive text classification (Baeza-Yates et al., 2011). We reinvigorate this old idea with modern techniques in text representation, retrieval, and data subset selection, giving rise to a practical and effective method for dataless text classification.

We evaluate the proposed approach through extensive experiments on a variety of datasets, including topical and sentiment classification tasks, multi-class and multi-label classification settings, and corpora from different genres. These experiments show that our approach often outperforms entailment-based methods by a large margin, enjoys fast prediction speed, and improves quickly if labeled documents are available for fine-tuning.

Our main contributions are as follows:

- We propose a simple and effective dataless text classification method that selects a document subset returned by dense retrieval models as pseudo-labels for classifier training.

- Extensive empirical experiments show that our method is more practically useful than the state-of-the-art textual entailment approaches. It enjoys higher accuracy, faster prediction speed, and can be readily improved even a small amount of labeled data are available.

2 Related Work

Dataless text classification (Chang et al., 2008) aims to classify text using a given set of class descriptions and no labeled data for training a model. Dataless text classification methods have two broad categories: classification-based (Chang et al., 2008; Druck et al., 2008; Wang and Domeniconi, 2009; Song and Roth, 2014; Yin et al., 2019; Chu et al., 2020a) and clustering-based (Barak et al., 2009; Chen et al., 2015; Li et al., 2016a, 2018; Li and Yang, 2018; Chu et al., 2020b). Classification-based methods use automatic algorithms to create machine-labeled data and construct a classifier that assigns a category to an input document. Clustering-based methods group documents (and class descriptions) by their similarity, and assign categories to each cluster. Our work focuses on the classification-based approach.

Several classic methods use explicit semantic analysis (ESA) (Gabrilovich et al., 2007) to represent documents and label descriptions in the same vector space of concepts, and then compute the cosine similarity between documents and labels. The label with the highest cosine similarity is assigned to the document as the classification result (Chang et al., 2008; Wang and Domeniconi, 2009; Song and Roth, 2014). These works emphasize that semantic representation of labels is as important as learning good representation of documents.

In previous works, dataless text classification also has many slightly different setups. For example, in zero-shot text classification, Yin et al.; Puri and Catanzaro proposed "label-fully-unseen" setting which directly computes document-label relatedness with a sentence-pair BERT model. The model is trained with large-scale texts naturally tagged with category information, such as Wikipedia. NATCAT takes a further step (Chu et al., 2020a). It combines various publicly available online corpora that come with natural categories, and trains a BERT or RoBERTa model (Devlin et al., 2018; Liu et al., 2019) to discriminate correct versus incorrect categories for a given document. These methods design automatic algorithms to create pseudo-labeled data from external resources to train a universal entailment model that can be applied to a wide spectrum of classification tasks.

It is easy to confuse "dataless text classification" with "zero-shot text classification" (Wang et al., 2019; Ye et al., 2020) and "weakly supervised text classification" (Meng et al., 2020a,b). Zero-shot text classification may still provide labeled data for part of the categories (label-partially-seen (Yin et al., 2019)), while dataless text classification does not assume labeled data for any category. Weakly supervised text classification assumes a large amount of unlabeled data are available for learning, while dataless text classification does not make this assumption – it can operate with few or no unlabeled data from the task domain.

3 Proposed Methods

In this section, we describe our proposed method for dataless text classification. We formulate the problem as follows. We are given a set of class descriptions $D = \{d_1, \cdots, d_j, \cdots, d_k\}$, each is a piece of short text (one or more words) describing a semantic class j in the label space $Y = \{1, \cdots, k\}$. We are given a set of unlabeled documents X and *zero* labeled documents in the task domain. As a natural scenario in practice, we also have access to vast amounts of external unlabeled documents U, $|U| >> |X|$. These external documents may come from Wikipedia, news corpora, and online social media, which may or may not share the same domain as the classification task in question. Our goal is to correctly assign label(s) from Y to (a subset of) unlabeled documents in either X or U as pseudo-labeled training data.

At a high level, our proposed method uses class descriptions in D as queries to retrieve pseudo-labeled documents from either task-specific unlabeled data X, or external unlabeled data U, or the two data sources combined. This gives us several variants of the method. We collectively name these variants **CLARET**, as they construct a classification model by leveraging a retrieval model. Below we describe our method in detail.

3.1 Dense Text Representation and Indexing

As a preparation step, we use a sentence representation model to convert all texts (class descriptions, task-specific unlabeled documents, and external unlabeled documents) into dense vectors in a semantic space. In principle, any dense text representation techniques can be used. We choose to use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) as it is proven to deliver good performance in various sentence-pair modeling and information retrieval tasks (Thakur et al., 2021).

Once these texts are converted into dense vectors, we build approximate nearest neighbor (ANN) indices for task-specific unlabeled documents and external documents to enable fast document retrieval. In principle, any ANN search techniques can be used. We choose to use FAISS (Johnson et al., 2017) for efficient similarity search with cosine similarity as the vector similarity metric. We also tested other metrics such as Euclidean distance but found negligible performance difference.

As SBERT is trained on a wide range of semantic similarity tasks (including textual entailment), the resulting document vectors inherit the knowledge from these tasks. Cosine similarity $\cos(x_1, x_2)$ between documents x_1 and x_2 approximates the probability that x_1 entails x_2 (or vice versa). In this sense, our method implicitly leverages the same type of knowledge of entailment-based models in a more efficiently computable manner.

3.2 Class-Relevant Document Retrieval

The first step of our method is to retrieve a pool of potentially relevant documents for each class, a subset of which will be pseudo-labeled in the next step. We propose three variants for this step.

Retrieving from task-specific unlabeled data. Oftentimes a classification task starts with task-specific data, but none of them are labeled yet. We use each class description as a search query to retrieve documents from task-specific unlabeled data. For class $j \in Y$, we rank documents in the unlabeled data X by their semantic similarity to the class description d_j and take the most similar n_1 documents $R_j = \{x_i\}_{i=1}^{n_1}$. Here, semantic similarity is computed using the vectors produced in Section 3.1. We call this variant $\textbf{CLARET}_{\textbf{task}}$.

$\text{CLARET}_{\text{task}}$ is most useful if abundant task-specific unlabeled data are available. However, sometimes even such data are few. For example, when mining documents related to an emerging event in a data stream, one may only collect a small number of documents about the new event since it just happened. In that case, task-specific data can be too scarce to retrieve from. To address this scarcity, we can instead retrieve from external data sources that contain vast amounts of unlabeled documents, some of which can also be semantically related to the current task. This is the next variant.

Retrieving from external data. We can retrieve class-relevant documents from external data when task-specific data is scarce. External data should come from as rich and diverse sources as possible to increase the chance of returning task-relevant documents. Thanks to approximate nearest neighbor search index, the retrieval step can be done efficiently against arbitrarily large external data. For class j, we retrieve n_2 most relevant documents from external data with respect to the class description d_j, $R_j = \{x_i\}_{i=1}^{n_2}$. We call this variant **CLARET$_{\text{external}}$**.

Retrieving from external data with a task-specific focus. We consider a third variant that combines the previous two. The idea is to enrich a class description with task-specific data before using it to retrieve external documents. For each class j, we first obtain a "seed set" of documents S_j using the same approach as CLARET$_{\text{task}}$ by fixing $n_1 = .1 \times |X|/k$. Then we use them to further retrieve external documents by treating each $x \in S_j$ as a query to retrieve its n_3 nearest neighbors $\Gamma(x)$ from external data. However, these documents may be close to a seed document because they share words unrelated to the theme of the class. To filter such noise, we preserve documents that appear in at least two seed documents' nearest neighborhoods. This gives class-relevant documents for class j: $R_j = \{e|\exists x_1, x_2 \in S_j, e \in \Gamma(x_1) \wedge e \in \Gamma(x_2)\}$. The hope is that R_j contains external documents that are semantically relevant and stylistically similar to task-specific data. We call this variant **CLARET$_{\text{task-external}}$**.

3.3 Pseudo-Labeled Subset Selection

A challenging problem remains: how many documents to retrieve and assign pseudo-labels (namely, how to set n_1, n_2, or n_3)? More generally, what is the optimal subset of retrieved documents that, if pseudo-labeled, will train a good classifier? Note that we cannot tune subset selection procedures on labeled data as such data is unavailable in a dataless setting! We propose a novel *unsupervised* subset selection procedure to address this problem.

Subset diversification. For CLARET$_{\text{task}}$, we create pseudo-labeled set $L_j = \{(x, j)|x \in R_j\}$ for each class j. For the two variants that use external data (CLARET$_{\text{external}}$ and CLARET$_{\text{task-external}}$), however, we further select a subset $L_j \subset R_j$ of size m to be pseudo-labeled as class j. The motivation is that documents retrieved from external data sources may contain (near-)duplicates. For example, many news outlets may cover the same story. Duplicated documents may lead to overfitting as they give too much emphasis on a few documents and reduce the overall diversity of pseudo-labeled training data. Indeed, previous works have shown that diverse training data improves learning performance (Wei et al., 2015). Here we apply facility location function to quantify the diversity of a subset (Krause and Golovin, 2014). The facility location function of any subset $L_j \subset R_j$ is defined as

$$g(L_j) = \sum_{x \in R_j} \max_{e \in L_j} s(x, e) \; . \qquad (1)$$

Here $s(\cdot, \cdot)$ is the cosine similarity between two dense document vectors. Intuitively, $g(L_j)$ computes the total cost for every element $x \in R_j$ to be "covered" by the most similar element $e \in L_j$. In our context, this translates into how well the subset L_j preserves the content of the larger set R_j. Although finding the optimal subset L_j that maximizes the submodular function $g(L_j)$ is NP-hard, a greedy algorithm gives an approximately optimal solution (Nemhauser et al., 1978). The algorithm sequentially adds the next element x to L_j with the maximum marginal gain $g(L_j \cup \{x\}) - g(L_j)$, until L_j reaches the desired size m.

Entropy maximization. We now determine the subset selection parameters θ. For CLARET$_{\text{task}}$, $\theta = \{n_1\}$. For CLARET$_{\text{external}}$, $\theta = \{n_2, m\}$. For CLARET$_{\text{task-external}}$, $\theta = \{n_3, m\}$. θ determines the pseudo-labeled set L_j for class j, which determines the full pseudo-labeled set $\cup_{j=1}^{k} L_j$, which in turn trains a classifier $f : X \to Y$. Below we use f_θ to emphasize that f depends on θ. f_θ induces a distribution over the label space Y when applied to the task-specific unlabeled data X: $\forall y \in Y$,

$$p(y|X, f_\theta) = \frac{\sum_{x \in X} \mathbf{1}\{f_\theta(x) = y\}}{|X|} \; . \qquad (2)$$

According to the maximum entropy principle (Jaynes, 1957), the distribution with maximum entropy shall be preferred since *no* labeled data are available as evidence to prefer other distributions. Following this principle, we seek for θ that maximizes the classification entropy:

$$H(\theta) = \sum_{y \in Y} -p(y|X, f_\theta) \log p(y|X, f_\theta) \; . \qquad (3)$$

Empirically, $H(\theta)$ correlates well (but not perfectly) with true performance of f_θ on labeled data even though it is an unsupervised metric (Appendix C.3), a phenomenon first observed in (Baram et al., 2004). As $H(\theta)$ is non-differentiable with respect to θ, we resort to grid search. It is sufficient to use a coarse grid to find sensible θ values (Section 4.3).

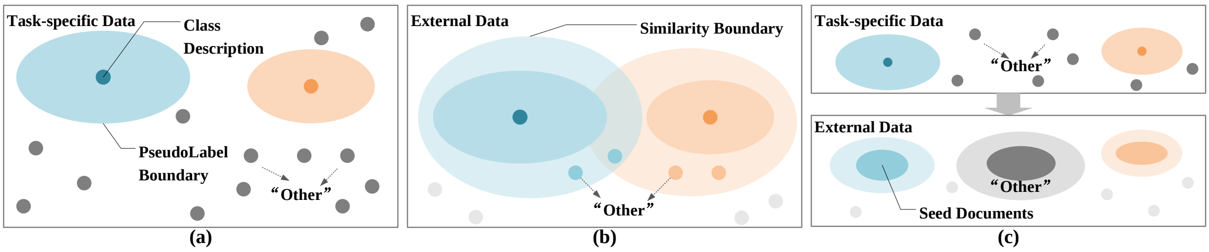


Figure 1: Handling the *Other* class. (a) CLARET$_{\text{task}}$: obtaining *Other* documents from task-specific data; (b) CLARET$_{\text{external}}$: obtaining *Other* documents between pseudo-label bounary and similarity boundary from external data; (c) CLARET$_{\text{task-external}}$: retrieving *Other* documents from external data using seeds from task-specific data.

3.4 Handling the *Other* Class

In some classification tasks, we have clearly defined categories and an *Other* category, such as an "other topic" category in topic classification or a "no emotion" category in emotion classification. We call clearly defined (non-*Other*) categories *named classes*. Using "other topic" or "no emotion" literally as the search query to retrieve pseudo-labeled documents is problematic because the *Other* class is to be interpreted with respect to named classes. We propose methods to handle the *Other* class for each variant above. The general idea is to pseudo-label documents that are far from any named class as the *Other* class. Without loss of generality, let the named classes be numbered from 1 to $k-1$ and the *Other* class be class k.

For CLARET$_{\text{task}}$, we select *Other* documents L_k from task-specific unlabeled data $O = X \backslash \cup_{j=1}^{k-1} L_j$. Our goal is to find a subset $L_k \subset O$ with size n_1 that is farthest from the descriptions of all named classes $D \backslash \{d_k\}$. We seek for the subset that *minimizes* the following function:

$$h(L_k) = \sum_{x \in L_k} \max_{e \in D \backslash \{d_k\}} s(x, e) \,. \tag{4}$$

This function is modular and can be efficiently minimized by selecting n_1 documents that have smallest $\max_{j=1}^{k-1} s(d_j, x)$ values from O (Figure 1a).

For CLARET$_{\text{external}}$, we first retrieve external data that are far from all named class descriptions but still relevant to the task: $O = \cup_{j=1}^{k-1} \{x | x \in U, 0 < s(x, d_j) < 0.1\}$. We then select $R_k \subset O$ with size n_2 by optimizing $h(R_k)$ (Eq. (4)), and then use the same diversity and entropy maximization procedure in Section 3.3 to select m documents in R_k and pseudo-label as *Other* (Figure 1b).

For CLARET$_{\text{task-external}}$, we first use the same procedure as CLARET$_{\text{task}}$ (Eq. (4)) to select task-specific seed documents for the *Other* category. This turns *Other* into another named class. We then retrieve and select pseudo-labels using the same procedure described in Sections 3.2 and 3.3 (Figure 1c).

4 Experiments

In this section, we evaluate our proposed methods and compare them with baseline models for dataless text classification. The comparison is not only in terms classification accuracy, but also label efficiency and inference speed.

4.1 External Document Repository

To cover various task domains, we combine five large-scale datasets as the external document repository. These datasets are freely available and frequently used in previous works as external resources. We keep these documents short (e.g. titles) as SBERT is well-trained on sentence pairs. We build a single index for all the external documents.

Microsoft News Dataset (MIND) (Wu et al., 2020) is collected from anonymized behavior logs of Microsoft News website. **Multi-Domain Sentiment Dataset (MDSD)** (Blitzer et al., 2007) contains product reviews for many product categories in Amazon. **Wikipedia-500K** (Bhatia et al., 2016) has over a million curator-generated category labels and each article often has more than one relevant labels. We select the first sentence of each article. **RealNews** (Zellers et al., 2019) is a large news corpus from Common Crawl. We randomly sample 2M titles from these 32M news. **S2ORC** (Lo et al., 2020) is a general corpus of scientific literature. We randomly select 100k papers from all 20 research fields and extract their titles.

4.2 Evaluation Datasets

We choose 10 text classification tasks in our experiments. Note that we do not use any data or labels from the training set, but only use unlabeled

Dataset	#Docs	#Sents/doc	#Words/doc
MIND	98,336	1	10.7
MDSD	821,250	7.3	137.5
Wikipedia	1,779,881	1	22.9
RealNews	2,000,000	1	9.6
S2ORC	2,000,000	1	10.9

Table 1: Statistics of external document datasets.

documents in the test set and the original class descriptions (see Appendix A).

Single label topic classification datasets. **Yahoo** (Zhang et al., 2015) consists of 10 categories of questions in online forums. **20Newsgroup** (Lang, 1995) is a collection of 20 topic newsgroup documents. **AGnews** (Zhang et al., 2015) contains 4 topical categories of news tities. **DBPedia** (Lehmann et al., 2015) contains titles, descriptions, and associated categories from DBpedia.

Single label sentiment classification datasets. **Yelp** (Zhang et al., 2015) is for sentiment analysis in Yelp reviews. **Emotion** (Oberländer and Klinger, 2018) was constructed by combining multiple public datasets where documents have emotion labels. **Amazon** (Zhang et al., 2015) is a binary sentiment classification dataset. **SST** (Socher et al., 2013) is a corpus extracted from movie reviews.

Multi label topical classification datasets. **Situation** (Zhang et al., 2015) is a event-type classification dataset originally designed for low-resource situation detection. **Comment** is created by Chu et al. and contains 28 classes.

Dataset	#Docs	#Classes	#Docs/class	#Words/doc
Single-label topic classification				
Yahoo	100k	10	10K	115.8
AGnews	7,600	4	1,900	48.8
20News	7,532	20	376	375.4
DBPedia	70k	14	5,000	58.7
Single-label sentiment classification				
Yelp	38k	2	19K	155.1
Emotion	16k	10	1,600	19.5
Amazon	400k	2	200K	95.7
SST-B	1,821	2	910.5	19.2
Multi-label topic classification				
Situation	3,525	12	380.2	44.0
Comment	1,287	28	90.7	13.8

Table 2: Statistics of evaluation datasets.

4.3 Compared Methods

We include two state-of-the-art methods for dataless text classification: *label-fully-unseen* 0SHOT-TC (Yin et al., 2019) and NATCAT (Chu et al., 2020a). These two methods both use readily available resources to train textual entailment models that can robustly handle a wide range of text classification tasks. To study the contribution of a dense retrieval model in our approach, we construct a baseline by replacing SBERT+FAISS with sparse text retrieval model (BM25).

Label-fully-unseen **0SHOT-TC** was first explored in (Yin et al., 2019). This setting pushes “zero-shot learning” to the extreme – no annotated data for any labels. It aims to classify documents without seeing any task-specific training data. They trained an entailment-based classifier on MNLI, FEVER and RTE datasets to predict a binary outcome. In the testing phase, they converted category descriptions into hypothesis in two ways, one is to prefix the label description with “it is related to”, the other is to use WordNet definition of the category label words in a hypothesis.

NATCAT (Chu et al., 2020a) proposed to use large-scale, naturally annotated data to train robust entailment-based text classification models. The authors induced document-category pairs from Wikipedia, Stack Exchange, and Reddit posts. Unlike label-fully-unseen 0SHOT-TC, NATCAT did not convert each category into a hypothesis, but directly connected the category and the document as a sentence-pair input.

BM25 retrieval. This baseline uses BM25 instead of SBERT+FAISS for document retrieval in CLARET$_{\text{task-external}}$. We build two inverted indices, one for task-specific data, the other for external data. Using class descriptions as queries, we use BM25 to retrieve n_1 task-specific documents and select 20 class-specific keywords using TF-IDF scores of words in retrieved documents. Then we use these class-specific keywords as queries to retrieve n_3 documents from the external data. Finally, we still use the facility function to filter m documents from the external data. The parameter settings (n_1, n_3, m) are the same as CLARET$_{\text{task-external}}$. Document indexing and BM25 document retrieval are implemented using the Python Whoosh library.

The three variants of **CLARET** we proposed. To select pseudo-labeled subsets that have maximum classification entropy, we searched parameters θ on the grids $n_1 = \{.1, .3, .5\} \times |X|/k$, $m = \{100, 300, 500\}$, $n_2 = \{2m, 5m, 10m\}$ and $n_3 = \{100, 200, 300\}$. The subset-induced RoBERTa classifier that achieved the maximum entropy was

Method	Single-label								Multi-label	
	Yahoo	AGnews	20News	DBPedia	Yelp	Emotion	Amazon	SST	Situation	Comment
Baseline Models										
BM25	39.6	69.7	31.1	68.6	49.5	13.2	52.0	52.2	14.0	15.1
0SHOT-TC (best)	43.8	-	-	-	-	24.7	-	-	**37.2**	-
0SHOT-TC (our)	24.9	67.8	19.0	58.0	71.0	21.1	78.3	68.6	20.1	22.3
NATCAT (best)	57.8	75.6	39.3	82.8	70.4	-	66.8	65.0	-	22.6
NATCAT (our)	48.6	74.9	44.8	85.3	50.1	10.7	50.8	50.5	27.4	22.0
CLARET										
Task	56.1	77.4	57.2	83.0	83.4	**28.4**	78.4	**85.5**	11.5	21.8
External	57.3	72.7	51.7	84.9	**87.9**	27.6	**89.5**	80.1	30.5	23.3
Task-External	**61.6**	**84.5**	**58.3**	**92.7**	86.5	27.1	86.2	84.1	**37.2**	**25.9**

Table 3: Dataless text classification performance on ten datasets (%). Each metric of CLARET is the average of 5 runs with different random seeds. The metrics are label ranking average precision (LRAP) for Comment, label-weighted F1 for Emotion and Situation and accuracy for other single-label classification tasks. The best reported results of label-fully-unseen 0SHOT-TC results from (Yin et al., 2019) and weakly supervised model NATCAT (Chu et al., 2020a) are included. We also report results of our re-implementation of 0SHOT-TC pre-trained on MNLI and NATCAT model pre-trained on Wikipedia. Both used RoBERTa as the entailment model. The best average performance in each column is highlighted in bold.

used. The optimizer is AdamW (Loshchilov and Hutter, 2017), learning rate is $2e^{-5}$, training batch size is 32 and the number of training epochs is 4.

We did not compare with the LOTClass model (Label-Name-Only Text Classification) (Meng et al., 2020a). LOTClass assumes that label words are mentioned somewhere in unlabeled documents, which is not guaranteed. For example, in Emotion and Yahoo datasets, some label words are not mentioned in any documents. Also, LOTClass does not deal with the *Other* class, which is present in Emotion and Situation datasets.

4.4 Performance Across Datasets

Table 3 summarizes classification performance of baseline methods and our three pseudo-labeling methods combined with RoBERTa classifier. Besides, we have stored our implementation as open source code in an anonymous Github repository[1].

These results show that variants of CLARET are able to achieve the highest performance on each task compared with baseline methods. Although the best pseudo-labeling strategy depends on specific tasks, it is clear that CLARET is overall a promising approach to dataless text classification. It performs the same as or sometimes much better than entailment models. Comparison of BM25 and CLARET variants shows that dense retrieval module (e.g., SBERT+FAISS) is essential in obtaining pseudo-labeled documents. (See Appendix C for supplementary performance analysis.)

[1]https://anonymous.4open.science/r/CLARET-6FD2

4.5 Prediction Speed Comparison

A big advantage of classification models over entailment models is the prediction speed. Classification models only need one forward pass to make a prediction for k categories, whereas entailment models need k forward passes. Table 4 compares prediction time of entailment models and $\text{CLARET}_{\text{task-external}}$ on the Yahoo dataset (100,000 documents). Our method is not only more accurate (Table 3) but also 5-7 times faster.

Method	Total Time	Per Document
0SHOT-TC	2162.4s	22ms
NATCAT	1485.8s	15ms
$\text{CLARET}_{\text{task-external}}$	**306.7s**	**3ms**

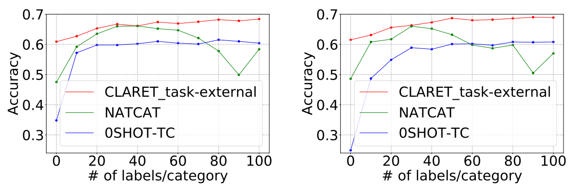
Table 4: Total testing time on Yahoo using *label-fully-unseen* 0SHOT-TC, NATCAT and $\text{CLARET}_{\text{task-external}}$. All methods used RoBERTa-base model.

Although entailment models are universal which only need to be trained once to be applied to any task, in order to obtain excellent results, a large amount of entailment data are required for pre-training. NATCAT uses three different data sources, a total of 10M training documents for pre-training. We measured the pre-training time using only Wikipedia data, which already took more than 50 hours. For 0SHOT-TC, since there is no author-released code for pre-training, we used MNLI data, batch size = 64, and 3 training epochs. It took about 2 hours. In our method, indexing external data repository took about 45 minutes. Taking the Yahoo dataset as an example, we measured the time to index the dataset, retrieve pseudo-labeled documents, select pseudo-label subsets and train a classifier using CLARET$_{\text{task-external}}$. The entire process took about 2 hours. Other datasets typically took less time as the Yahoo dataset has many categories and each retrieves many class-relevant documents. Therefore, although our methods take time to train classifiers for new tasks, the cost of training time can be amortized by the saving of prediction time in the long run compared to entailment-based models.

4.6 Learning Curve Comparison

Practioners may wish to further improve a dataless classification model as its initial performance can be far from optimal. We therefore ask the question: if a small amount of training data becomes available, how fast can a dataless model improve?

To verify our hypothesis that with continuous increase of training data, a classification model will improve faster than an entailment model, we present a learning curve analysis using Yahoo dataset. We compare entailment models *label-fully-unseen* 0SHOT-TC (Yin et al., 2019) pre-trained on MNLI, NATCAT (Chu et al., 2020a) pretrained on Wikipedia, and our classification model trained on CLARET$_{\text{task-external}}$ pseudo-labels. We use the same set of labeled documents with increasing sizes, the learning rate is $5e^{-5}$ and training epochs is 4 to fine-tune each of the three models.



(a) BERT (b) RoBERTa

Figure 2: Learning curves of CLARET$_{\text{task-external}}$ and two entailment approaches when fine-tuned on increasing amount of training data from the Yahoo dataset.

The learning curves in Figure 2 show that compared with entailment models, the advantage of the classification model is not only in the initial high performance. We see from the learning curves that when each category has a certain amount of training data, the classification model shows the fastest performance gain. In contrast, the performance of entailment models flattens and even drops. This demonstrates that applying entailment models on dataless classification tasks has certain limitations. In fact, textual entailment is a much harder problem than text classification, as the former aims to learn pairwise dependencies between all words in the premise (document) and all words in the hypothesis (class description), while the latter aims to associate a document to a categorical variable. Therefore, an entailment approach to classification is indirect and label-inefficient.

4.7 Discussion

In Table 3, we not only report the results of our baseline models reported in previous works, but also the results implemented by ourselves. Here we make a special remark on the Situation and Emotion datasets: they both contain the *Other* class. For Situation this category is "out-of-domain" and for Emotion it is "no emotion". We handled the *Other* classes using the approach in Section 3.4.

The three proposed strategies all have their own advantages. The CLARET$_{\text{task-external}}$ strategy is suitable for topic classification tasks, whether it is single-label or multi-label. It chooses a small set of test documents as seeds and expand the document search on vast external data sources. For sentiment classification tasks, CLARET$_{\text{task-external}}$ does not obtain the best results but still outperforms the entailment model. CLARET$_{\text{task}}$ and CLARET$_{\text{external}}$ are suitable for sentiment classification tasks. CLARET$_{\text{task}}$ performs better on smaller datasets (Emotion, SST), while CLARET$_{\text{external}}$ performs better on Amazon and Yelp datasets. The crucial reason is that the Multi-Domain Sentiment Dataset in our external data consists of Amazon reviews data. Though Emotion is a sentiment classification task, its documents come from Twitter. Even though documents from the two data sets may express similar emotions, the transferable knowledge from Amazon reviews to tweets is limited due to different text styles. Therefore, CLARET$_{\text{task}}$ can achieve good results on Emotion and SST datasets.

5 Conclusion

We proposed a dataless text classification method CLARET which constructs a classification model by leveraging a dense retrieval model. Extensive experiments show that the proposed method is not only able to achieve excellent dataless classification performance, but also enjoys fast prediction speed and can be effectively improved when labeled training data become available, making it readily applicable in practical classification tasks.

References

Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. 2011. *8.3.2 Naive Text Classification*, chapter 8. New York: ACM Press; Harlow, England: Addison-Wesley,.

Libby Barak, Ido Dagan, and Eyal Shnarch. 2009. Text categorization from category name via lexical reference. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 33–36.

Yoram Baram, Ran El Yaniv, and Kobi Luz. 2004. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar):255–291.

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

John Blitzer, Mark Dredze, Fernando Pereira, and Bollywood Biographies. 2007. Boom-boxes and blenders: domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Pereira*, volume 447.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. Dataless text classification with descriptive lda. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Zewei Chu, Karl Stratos, and Kevin Gimpel. 2020a. Natcat: Weakly supervised text classification with naturally annotated datasets. *arXiv preprint arXiv:2009.14335*.

Zewei Chu, Karl Stratos, and Kevin Gimpel. 2020b. Unsupervised label refinement improves dataless text classification. *arXiv preprint arXiv:2012.04194*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Edwin T Jaynes. 1957. Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Andreas Krause and Daniel Golovin. 2014. Submodular function maximization. *Tractability*, 3:71–104.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016a. Effective document labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 85–94.

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. 2018. Dataless text classification: A topic modeling approach with document manifold. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 973–982.

Ximing Li and Bo Yang. 2018. A pseudo label based dataless naive bayes algorithm for text classification with seed words. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1908–1917.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. 2016b. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2678–2688.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Tingting Ma, Jin-ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020a. Text classification using label names only: A language model self-training approach. In *EMNLP (1)*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020b. Weakly-supervised text classification using label names only. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

J. J. Rocchio. 1965. Relevance feedback in information retrieval, report no. *ISR-9 to the National Science Foundation, The Computation Laboratory of Harvard University, to appear August*.

Jacob M Schreiber, Jeffrey A Bilmes, and William Stafford Noble. 2020. apricot: Submodular selection for data summarization in python. *J. Mach. Learn. Res.*, 21:161–1.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Twenty-eighth AAAI conference on artificial intelligence*.

Yangqiu Song, Shyam Upadhyay, Haoruo Peng, and Dan Roth. 2016. Cross-lingual dataless classification for many languages. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2901–2907.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Pu Wang and Carlotta Domeniconi. 2009. Towards a universal text classifier: Transfer learning using encyclopedic knowledge. In *2009 IEEE International Conference on Data Mining Workshops*, pages 435–440. IEEE.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963. PMLR.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

A Class Descriptions in Evaluation Datasets

We list the class descriptions of the datasets we used for evaluation as follows. These texts are used as to compute SBERT vector representations. Note that some class descriptions are very abstract: "positive" and "negative" for sentiment classification datasets (Yelp, Amazon, SST-B).

Yahoo: Society&Culture; Science&Mathematics; Health, Education&Reference; Computers&Internet; Sports; Business&Finance; Entertainment&Music; Family&Relationships; Politics&Government.

AGnews: politics; sports; business; technology.

20Newsgroup: atheist atheism; computer graphics; computer OS microsoft windows miscellaneous; computer system IBM PC hardware; computer system Mac hardware; computer windows xp; miscellaneous for sale; recreational automobile; recreational motorcycles; recreational sport baseball; recreational sport hockey; science cryptography; science electronics; science medicine; science space; society religion christian; talk politics guns; talk politics middle East; talk politics miscellaneous; talk religion miscellaneous.

DBPedia: Company; Educational Institution; Artist; Athlete; Office Holder; Mean Of Transportation; Building; Natural Place; Village; Animal; Plant; Album; Film; Written Work.

Yelp: positive; negative.

Amazon: positive; negative.

SST-B: positive; negative.

Emotion: anger; sadness; surprise; love; fear; disgust; guilt; shame; joy; no emotion.

Situation: utilities energy or sanitation; water supply; search/rescue; medical assistance; infrastructure; shelter; evacuation; regime change; food supply; crime violence; terrorism; out-of-domain.

Comment: team war; injury; sentiment; player humor; player praise; statistic; sentiment positive; communication; game praise; feeling; teasing; referee; audience; coach negative; sentiment negative; player; team caveat; game expertise; player criticize; commercial; coach positive; play; coach; commentary; referee positive; game observation; referee negative; team.

B Implementation Details

We implement the models with the same PyTorch framework and run the model on NVIDIA GeForce RTX 3090. Below, we summarize the implementation details that are key for reproducing results.

We use "paraphrase-MiniLM-L6-v2" as the base model for SBERT to obtain the sentence embeddings and the dimension of embedding vectors is 384. And we use FAISS to retrieve external documents which works with inner product to compute cosine similarity. The number of clusters is set to 512 and 3 clusters are explored at search time. We implemented facility location subset selection using the Apricot library (Schreiber et al., 2020), which provides cosine as a similarity measure and a lazy greedy optimizer as a solver. We train BERT and RoBERTa on the task datasets for dataless text classification. In our experiments, we use BERT-base-uncased (110M parameters) and RoBERTa-base (110M parameters).

C Additional Performance Analysis

C.1 BERT-based Classifier Performance

We have reported the results based on RoBERTa as our main result. Here we show the classification performance of baseline methods and our three pseudo-labeling methods all based on BERT classifier in Table 5. In most cases, we found that the performance of RoBERTa model is better than BERT. This may be because compared with BERT's use of Wikipedia and books the training data of RoBERTa comes from web text which is more diverse.

C.2 Supervised Classification Performance

We present the performance training with all the labeled data based on BERT and RoBERTa in Table 6. Here, we want to note that the Comment dataset is a provided by NATCAT(Chu et al., 2020a) for dataless classification, and it has test set only. So we randomly split 80% data from the official test set as training data and the other 20% data for test.

C.3 Relation Between Entropy and Accuracy

In order to verify the relationship between entropy and classification accuracy, we compared the trends of entropy and predicate accuracy under different parameter settings. Figure 3 shows the relation between the entropy and accuracy in Yahoo, SST, and Situation datasets. From Figure 3 we can see that with different parameters, the trends of entropy and accuracy are often (but not perfectly) correlated. It shows that the empirical classification entropy on unlabeled data is an effective unsupervised metric to guide the selection of pseudo-labeled subset.

Method	Single-label								Multi-label	
	Yahoo	AGnews	20News	DBPedia	Yelp	Emotion	Amazon	SST	Situation	Comment
Baseline Models										
BM25	41.6	69.8	27.8	59.2	54.9	11.1	49.8	51.8	13.5	14.0
0SHOT-TC (our)	34.8	53.8	22.2	53.8	73.4	21.7	76.0	71.7	16.2	22.6
NATCAT (our)	47.5	77.9	40	88.2	73.9	22.2	72.9	65.8	26.5	23.5
CLARET										
Task	55.9	77.5	57.2	82.2	82.9	**28.1**	78	82.4	11.1	17.2
External	56.7	74.6	49.9	86.1	83.3	26.6	**83.9**	79.5	28.1	21.1
Task-External	**60.7**	**82.5**	**57.5**	**93.0**	**83.9**	26.8	80.3	**83.1**	**35.2**	**23.9**

Table 5: Dataless text classification performance in ten datasets (%) based on BERT classifier. The best average performance in each column is in bold.

Method	Single-label								Multi-label	
	Yahoo	AGnews	20News	DBPedia	Yelp	Emotion	Amazon	SST	Situation	Comment
BERT	74.2	94.7	72.8	99.3	97.4	36.9	94.7	93.5	50.9	32.6
RoBERTa	75.1	95.4	73.5	99.3	97.5	37.8	97.4	95.8	58.4	33.8

Table 6: Dataless text classification performance in ten datasets (%) based on BERT and RoBERTa classifier training with full label-data.

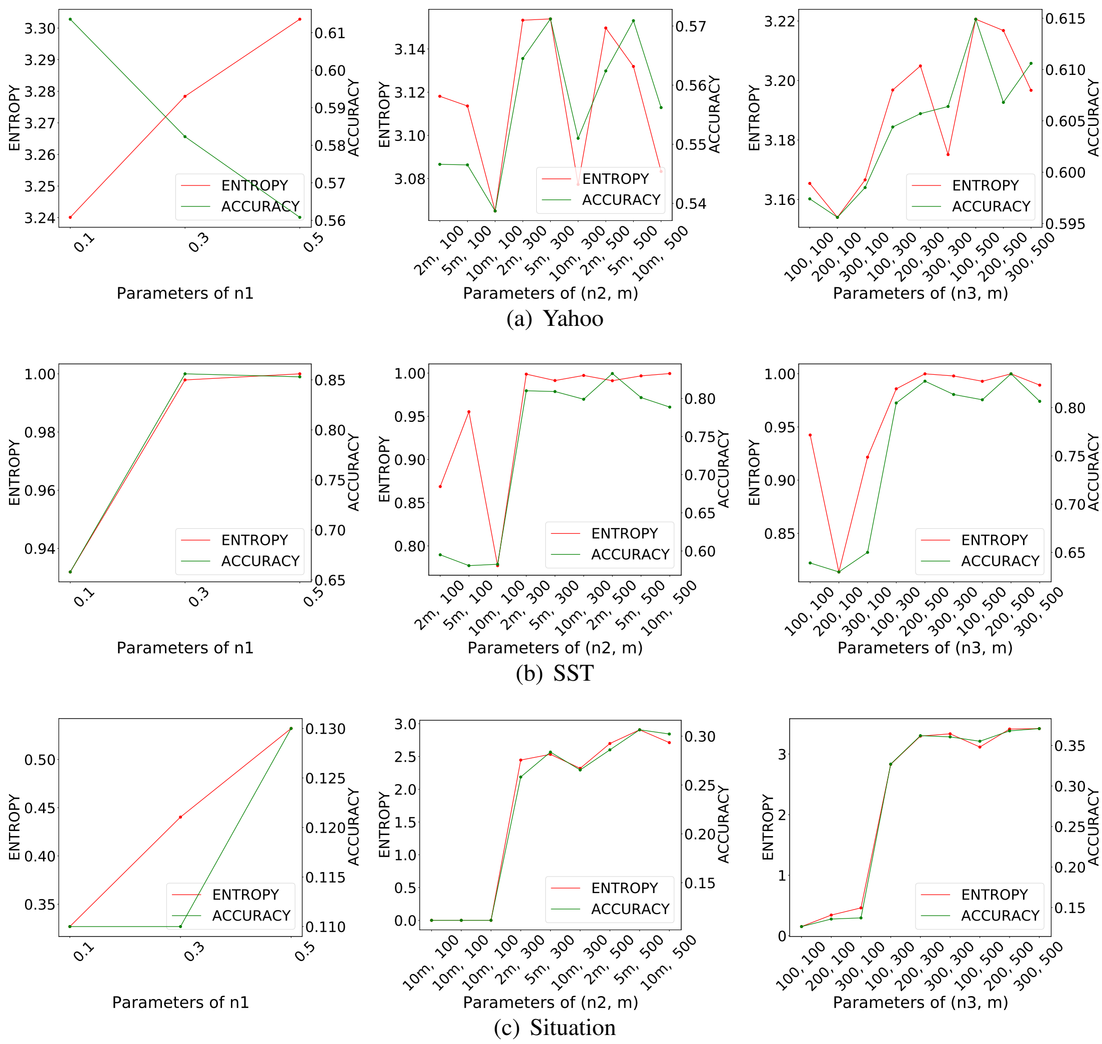


Figure 3: Relation between entropy and accuracy.