Text Classification Under Class Distribution Shift: A Survey

Adriana-Valentina Costache[1,*], **Silviu-Florin Gheorghe**[1,*], **Eduard Poesina**[1,*], **Paul Irofti**[1], **Radu Tudor Ionescu**[1,◇]
[1]Department of Computer Science, University of Bucharest, Romania
[*]Equal contribution. [◇]raducu.ionescu@gmail.com

Abstract

The basic underlying assumption of machine learning (ML) models is that the training and test data are sampled from the same distribution. However, in daily practice, this assumption is often broken, i.e. the distribution of the test data changes over time, which hinders the application of conventional ML models. One domain where the distribution shift naturally occurs is text classification, since people always find new topics to discuss. To this end, we survey research articles studying open-set text classification and related tasks. We divide the methods in this area based on the constraints that define the kind of distribution shift and the corresponding problem formulation, i.e. learning with the Universum, zero-shot learning, and open-set learning. We next discuss the predominant mitigation approaches for each problem setup. We further identify several future work directions, aiming to push the boundaries beyond the state of the art. Finally, we explain how continual learning can solve many of the issues caused by the shifting class distribution. We maintain a list of relevant papers at https://github.com/Eduard6421/Open-Set-Survey.

1 Introduction

The primary assumption of any machine learning (ML) model is that the data is independent and identically distributed (IID) (Vapnik, 1995). In learning theory, the IID assumption plays a critical role, as it guarantees the generalization capacity of models, provided that sufficient training data is available, and that the hypothesis class is not too large. However, this assumption does not always hold in daily practice. One such example is text categorization by topic, where new topics naturally emerge as the interest of people changes over time. For example, journalists started publishing news articles about the COVID-19 pandemic only after its outbreak began in December 2019[1]. Hence, if a classifier is trained in a closed-world setup (a scenario where it is presumed that every instance encountered by the model belongs to a class that is present in the training data), there is a high risk that the respective model will encounter instances from classes that were not present in the training data, during operation. Such instances will be misclassified as belonging to one of the training classes, degrading the overall performance of the ML system, and implicitly reducing user satisfaction.

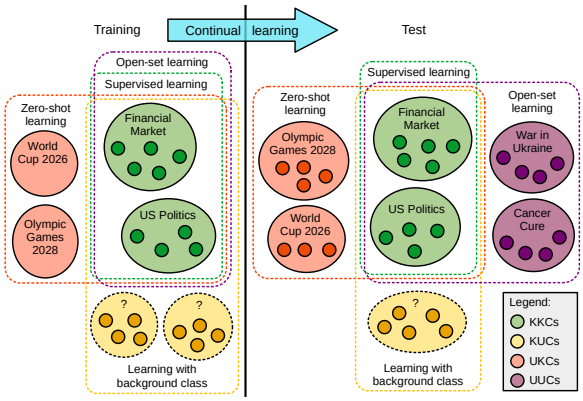

Figure 1: An illustration of text classification by topic under class distribution shift, resulting in four class categories and associated learning problems. Known Known Classes (KKCs) correspond to standard supervised learning, Known Unknown Classes (KUCs) to learning with background class, Unknown Known Classes (UKCs) to zero-shot learning, and Unknown Unknown Classes (UUCs) to open-set learning. Continual learning provides viable solutions to mitigate class distribution shift. Best viewed in color.

A large body of works, e.g. (Geng et al., 2021; Salehi et al., 2022; Vaze et al., 2021; Yang et al., 2024), studies the adaptation and application of ML models when the class distribution changes from training time to inference time. One of the first articles to highlight this problem is that of Scheirer et al. (2013). The authors define the *open-set risk* as the sum of the empirical risk and the open-space risk, the latter being defined as the risk of misclassifying instances of unknown classes as belonging to a known class. Over time, researchers attempted to minimize the open-set risk by addressing a broad range of formulations of the class distribution shift

[1]https://en.wikipedia.org/wiki/COVID-19_pandemic

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4045–4060

Category	Definition	Problem Formulation	Representative Studies
KKCs	Classes for which we have labeled training data.	Standard supervised learning	(out of scope)
KUCs	Classes with available training examples, but for which there are no class labels.	Classification with background / Universum class	(Dhamija et al., 2018; Hendrycks et al., 2019; Lee et al., 2018; Liu et al., 2020)
UKCs	Classes that are known to exist, but for which we have no examples during training.	Zero-shot learning	(Yin et al., 2019; Gera et al., 2022; Meng et al., 2020, 2022; Sanh et al., 2022; Zhang et al., 2024)
UUCs	Classes whose existence is unknown during training and for which there is no training data.	Open-set learning (and discovery)	(Chen et al., 2023; Walkowiak et al., 2020; Chen et al., 2024; Kim et al., 2022; Walkowiak et al., 2019a)

Table 1: Types of classes that can naturally emerge in classification under class distribution shift. For each class category of interest, we provide the corresponding problem formulation as well as a set of representative works from the NLP or ML domains.

problem. These formulations can be mainly structured according to the class categories proposed by Geng et al. (2021).

We present the class categories and the corresponding problem formulations in Figure 1. We further define the class categories and provide representative studies for each problem formulation in Table 1. KKCs lead to the conventional supervised learning paradigm, which is well-studied and results often surpass human-level performance (Cozma et al., 2018; Tedeschi et al., 2023). Frameworks that deal with KUCs typically use a background (a.k.a. Universum) class (Dhamija et al., 2018; Weston et al., 2006), essentially extending the supervised learning setup with a new category where all the examples that do not belong to any of the KKCs are placed. The zero-shot learning paradigm (Yin et al., 2019; Chaudhary et al., 2024; Pourpanah et al., 2023) aims to deal with UKCs, namely with classes that are known in advance (at training time), but for which there are no data samples. Zero-shot learning aims to classify data samples into UKCs, but the main limitation is that the set of UKCs is fixed. As a consequence, dealing with UUCs is not possible in the zero-shot learning setup. Open-set learning (Geng et al., 2021; Scheirer et al., 2013) aims to partially address this challenge by identifying UUCs without having prior information about such classes during training. Sometimes, UUCs are identified via a two-stage pipeline that combines an outlier detection (a.k.a. novelty detection) model (Chen et al., 2023; Walkowiak et al., 2020, 2018) and a supervised model trained on KKCs. However, unlike zero-shot learning methods, open-set methods do not aim to classify the data samples into UUCs. A more comprehensive setup is proposed in (Zheng et al., 2022), where the authors aim not only to detect samples belonging to UUCs, but also to classify them. This framework, called *open-set learning and discovery*, can be seen as a generalization of zero-shot learning, where the set of UKCs can be updated during inference, therefore transforming it into a set of UUCs.

To date, open-set classification and the related tasks discussed above were mostly studied in the vision domain, where various tasks have been explored, such as object recognition (Vaze et al., 2021; Scheirer et al., 2013; Kong and Ramanan, 2021), semantic segmentation (Cen et al., 2021; Oliveira et al., 2021), object detection (Zheng et al., 2022; Dhamija et al., 2020; Liu et al., 2024), and video anomaly detection (Acsintoae et al., 2022; Wu et al., 2024), among others. Comparatively less attention has been dedicated to this family of tasks in the text domain (Chen et al., 2024; Kim et al., 2022; Fei and Liu, 2016). However, class distribution shift is a prevalent phenomenon in text classification. Aside from the example provided earlier about the changing topics over time, there are many other natural language processing (NLP) tasks where the class distribution can shift. For example, in authorship identification, new authors can emerge over time, so an authorship identification model needs to update the list of potential authors. Another task affected by class distribution shift is intent detection in conversational AI. For instance, a chat bot trained to recognize intents such as "book flight" or "check weather" may encounter new intents such as "play music" at test time. Similarly, in named entity recognition, a system trained to recognize entities such as "person" and "location" might face new entity types such as "event" or "disease" during operation. Despite the

clear likelihood of encountering new classes during inference in various NLP tasks, to the best of our knowledge, there is no survey on open-set text classification. To this end, we review articles that address the class distribution shift in the text domain. We further propose future work directions to address the challenges of open-set learning and discovery in NLP tasks, most of which stem from the continual learning paradigm (Wang et al., 2024).

It is perhaps important to note that class distribution shift is a particular kind of dataset shift. The dataset shift problem was studied by Moreno-Torres et al. (2012), who meticulously categorize the various types of dataset shift. However, the authors did not specifically include class distribution shift as an interesting scenario, but they recognize the existence of dataset shift scenarios that *"are so hard that we currently consider them impossible to solve"* (Moreno-Torres et al., 2012). Since 2012, the state of research in machine learning has changed significantly, and many problems that were thought as being impossible to solve are now actively studied. For instance, there are several recent surveys, e.g. (Geng et al., 2021; Parmar et al., 2023; Zhu et al., 2024), that focus on the open-world learning problem, where new classes can emerge during testing. Yet, most of the studies covered by these surveys come from computer vision and image processing. To the best of our knowledge, we are the first to focus specifically on class distribution shift in natural language processing.

In summary, our contribution is twofold:

- We provide a literature review of open-set text classification and related tasks, including zero-shot text classification and learning with the Universum class.
- We propose several future work directions for open-set learning and discovery, a task that has not been extensively explored in NLP.

2 Notations and Definitions

Notations. We use the following notations to define the various task formulations covered in our survey. Let $x \in \mathcal{X}$ represent a text sample from the data space $\mathcal{X}$, and $y \in \mathcal{Y}$ a label from the label space $\mathcal{Y}$. The label space is divided into four disjoint subsets denoted as known known classes (C^{kk}), known unknown classes (C^{ku}), unknown known classes (C^{uk}), and unknown unknown classes (C^{uu}), such that $C^{kk} \cup C^{ku} \cup C^{uk} \cup C^{kk} = \mathcal{Y}$ and the intersection between any two subsets is the empty set, e.g. $C^{kk} \cap C^{uk} = \emptyset$.

Learning with background class. Learning with background / Universum class aims to detect samples from C^{ku}, while having samples from both C^{kk} and C^{ku}. Formally, the training data is defined as $\mathcal{D} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{kk} \cup C^{ku}\}$, while the test data is defined as $\mathcal{T} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{kk} \cup C^{ku}\}$. In this setup, a classifier is defined as $h : \mathcal{X} \to C^{kk} \cup \{\text{background}\}$, i.e. for any sample that belongs to C^{ku}, the model should label the respective sample as background.

Zero-shot text classification. Zero-shot learning aims to recognize samples from C^{uk} using knowledge transferred from C^{kk}. Formally, the training data is defined as $\mathcal{D} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{kk}\}$, while the test data is defined as set $\mathcal{T} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{uk}\}$. Knowledge is usually transferred via an auxiliary semantic space A, e.g. word embeddings or attribute vectors, such that $f : \mathcal{X} \to A$ and $g : A \to \mathcal{Y}$. The mapping f projects features into the semantic space, while g maps semantic representations to class labels.

Open-set text classification and discovery. Open-set learning aims to classify samples from C^{kk}, while rejecting samples from C^{uu} encountered during inference. The training set is defined as $\mathcal{D} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{kk}\}$, while the test set is given by $\mathcal{T} = \{(x, y) \mid x \in \mathcal{X}, y \in C^{kk} \cup C^{uu}\}$. An open-set learning model is defined as $h : \mathcal{X} \to C^{kk} \cup \{\text{unknown}\}$, i.e. for any sample that belongs to C^{uu}, the model should label the respective sample as unknown. In the *open-set learning and discovery* setup, the model is defined as $h : \mathcal{X} \to C^{kk} \cup C^{uu}$, i.e. the model must be able to classify any data sample, regardless of the fact that the sample belongs to C^{kk} or C^{uu}. To classify samples into C^{uu}, classes belonging to C^{uu} must be discovered during inference.

3 Optimization Objectives

Starting from a model whose parameters θ are obtained by training on a given dataset $\mathcal{X}$ with labels $\mathcal{Y}$, our general optimization problem is:

$$\begin{aligned} \theta^\star &= \arg\min_\theta \mathcal{L}\left(\theta, \mathcal{X}, \mathcal{Y}\right) \\ &= \arg\min_\theta \mathbb{E}_{(x,y)\sim(\mathcal{X},\mathcal{Y})}\left[f_\theta(x, y)\right]. \end{aligned} \tag{1}$$

We recover below the particular learning frameworks by giving specific loss formulations w.r.t. the class types defined in Table 1. The supervised cross-entropy on KKCs is:

$$\mathcal{L}_{\text{KKC}}(\theta, \mathcal{X}, C^{kk}) = \mathbb{E}_{(x,y)\sim(\mathcal{X},C^{kk})}[-\log p_\theta(y \mid x)]. \tag{2}$$

The binary cross-entropy on KKCs vs. KUCs is:

$$\mathcal{L}_{\text{KUC}}(\theta, \mathcal{X}, C^{kk} \cup C^{ku}) = \mathbb{E}_{(x,y)\sim(\mathcal{X}, C^{kk}\cup C^{ku})} \\ [-y \cdot \log f_\theta(x) + (1-y) \cdot \log(1 - f_\theta(x))]. \quad (3)$$

The contrastive learning objective on UKCs is:

$$\mathcal{L}_{\text{UKC}}(\theta, \mathcal{X}, C^{uk}) = \mathbb{E}_{(x,y)\sim(\mathcal{X}, C^{uk})} \\ \left[-\log \frac{\exp\left(\text{sim}(f_\theta(x), s_y)\right)}{\sum_{c \in C^{uk}} \exp\left(\text{sim}(f_\theta(x), s_c)\right)}\right], \quad (4)$$

where s_y is a prototype vector (embedding) for class y, which can be obtained by passing the class names through the model f_θ, and sim is a similarity measure (e.g. the cosine similarity). Finally, the supervised cross-entropy on UUCs is:

$$\mathcal{L}_{\text{UUC}}(\theta, \mathcal{X}, C^{uu}) = \mathbb{E}_{(x,y)\sim(\mathcal{X}, C^{uu})}[-\log p_\theta(y \mid x)], \quad (5)$$

and the label set is computed as $C^{uu} = \{y \mid y = g_\omega(x'), \forall x' \in \mathcal{I}(h_\phi(x))\}$, where h_ϕ is an outlier detection method (e.g. Isolation Forrest) and $\mathcal{I}$ is an unsupervised outlier indicator function that selects outlier samples x', which are further given as input to a clustering method g_ω (e.g. k-means).

Further, we can integrate all learning frameworks into a joint optimization objective:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{KKC}} + \alpha \cdot \mathcal{L}_{\text{KUC}} + \beta \cdot \mathcal{L}_{\text{UKC}} + \gamma \cdot \mathcal{L}_{\text{UUC}}, \quad (6)$$

where α, β and γ are scalar values that can be used to control the optimization within a particular learning framework, as follows:

- Learning with background class: $\alpha > 0$, $\beta = 0$, $\gamma = 0$.
- Zero-shot learning: $\alpha = 0$, $\beta > 0$, $\gamma = 0$.
- Open-set learning and discovery: $\alpha = 0$, $\beta = 0$, $\gamma > 0$.

4 Learning with Background Class

The problem of learning with a background class has been addressed through proxy tasks such as out-of-distribution (OOD) detection or outlier rejection (Dhamija et al., 2018). These approaches improve the reliability of the model by managing uncertainty and rejecting inputs outside the distribution of KKCs. Research in this area has been predominantly driven by computer vision studies, whereas text classification remains comparatively underexplored. However, many techniques can be easily adapted to NLP tasks by replacing image-based feature extractors with robust language encoders.

A straightforward approach to detect background samples involves training a classifier to refine the decision boundary between the original dataset and a curated background dataset. This technique is commonly employed in computer vision tasks (Mullapudi et al., 2021), leveraging the abundance of background images available, e.g. iNaturalist-BG, Places-BG, etc.

Another common approach is to employ custom regularization techniques for background-class learning. For Support Vector Machines (SVMs), Weston et al. (2006) introduce a penalty term that incorporates the distance between the decision boundary and background-class (Universum-class) data points. Their method is based on the principle that Universum-class samples should lie near the decision boundary, reflecting real-world uncertainty and ambiguity in classification. This regularization technique enforces a decision-making process that accounts for inherent uncertainty in the data. Expanding this idea, Zhou et al. (2023) propose a learning method that distinctively treats the classes of interest and the Universum class. They introduce closed decision boundaries for the classes of interest and define the space outside these boundaries as belonging to the Universum. To determine whether a sample belongs to a class of interest or to the Universum class, the authors propose a probability estimation based on inter-class rules. This approach assesses whether a sample lies predominantly within the closed decision boundary associated with a class of interest. If the sample does not clearly fit into one of the classes, it is considered to belong to the Universum. Dhamija et al. (2018) introduce a more generic regularization based on two loss functions. The Entropic Open-Set Loss aims to reduce entropy for KKCs, while maximizing entropy for KUCs, ensuring that the model treats unfamiliar data with the highest possible uncertainty. The Objectosphere Loss strengthens the separation between KKCs and KUCs by increasing the feature magnitudes for KKC samples, while reducing them for KUCs. These loss functions create a more representative feature space, improving the ability of the model to reject out-of-distribution inputs.

Developing novel loss functions represents a promising line of work, considered by other researchers as well. For instance, Liu et al. (2020) propose energy-based models as a viable solution for background sample detection, where the main contribution is a loss function designed to include both a standard classification loss and a model-specific penalty term. This additional term penalizes high-energy representations of KKC samples, while suppressing low-energy assignments

for KUC samples, ensuring a clear energy-based separation between the two categories. Hendrycks et al. (2019) introduce the concept of outlier exposure, which involves training models on an auxiliary dataset containing OOD samples. Their objective function combines the original classification loss with an additional penalty that makes the model assign low confidence to out-of-distribution inputs. For softmax-based classification tasks, this approach encourages predictions for outliers to follow a uniform distribution. In density estimation tasks, the outlier exposure loss often incorporates a margin ranking objective, ensuring that in-distribution samples consistently receive higher log probabilities than outliers. Lee et al. (2018) propose a method to train neural networks for better OOD detection, while maintaining their classification accuracy. The contributions refer to introducing a confidence loss, leveraging a generative adversarial network (GAN) (Goodfellow et al., 2014) for OOD sample generation, and combining these into a joint training framework. The confidence loss combines the standard cross-entropy loss with a KL-divergence term penalizing confident predictions for OOD samples, which are compared with a uniform distribution. The OOD samples are generated by a GAN, such that the samples lie close to the boundary of the in-distribution data. The authors train the classifier and the GAN jointly. A complementary direction is explored by Hu and Khan (2021), who develop a loss function specifically designed to calibrate uncertainty in both in-distribution and OOD settings. Their framework employs evidential neural networks and adds two regularization terms: one penalizes high uncertainty for in-distribution samples, and the other rewards maintaining high uncertainty OOD samples. By explicitly modeling and balancing predictive confidence, their approach improves classification reliability under class distribution shift.

5 Zero-Shot Text Classification

When no examples from the target classes are available, various alternative tasks can be leveraged, with or without additional training, to perform zero-shot text classification. For instance, Yin et al. (2019) propose using the entailment task as a zero-shot text classification task. To decide if a text x should be classified into a certain class, e.g. "politics", one can ask if x entails "The previous text is about politics". This procedure is applied for each of the possible classes, and the one with the best confidence is considered the correct one. Using this technique, any model capable of solving the entailment problem can implicitly be used to perform zero-shot text classification. The zero-shot approach based on entailment (Yin et al., 2019) represents the basis for how large language models (LLMs) are used via prompting to solve various downstream tasks. LLMs (Bommasani et al., 2021; Yang et al., 2025; Zhou et al., 2024) are usually pre-trained on a very large corpus of unlabeled data, using a variety of self-supervised tasks, e.g. next token prediction, sentence order prediction, etc. During pre-training, LLMs learn general facts, such as language structure, making them ideal for task adaptation. LLMs are subsequently used in different downstream tasks, which often involve zero-shot setups. There are two main methodologies for adapting LLMs to a target task: *fine-tuning* and *instruction tuning*.

Fine-tuning involves adapting a model to a task by performing an extra training step using task-specific data. Instruction tuning, defined as fine-tuning language models on a collection of datasets described via instructions, can be seen as a generalization over the framework proposed by Yin et al. (2019). Instruction tuning can lead to better performance on unseen tasks, if the target task is described via simple instructions provided in natural language. Models trained with instruction tuning can perform various tasks, including zero-shot classification, as shown by Zhang et al. (2024). This capability is achieved by training language models to respond to simple natural language commands, with some degree of generality. For instance, Wei et al. (2022) and Sanh et al. (2022) explore similar ideas, dividing the tasks for which datasets are available into separate clusters, each containing multiple datasets. Fine-tuning via instructions on any of the clusters leads to performance gains when the model is tested on tasks from the other clusters. Xu et al. (2022) continue this line of work by increasing the number of tasks used for pre-training from a few dozen to over 1000, showing that increasing the number of tasks is a good alternative to increasing the model size.

Another possible approach for zero-shot classification is to use an LLM to generate a synthetic dataset $\mathcal{S}$, tailored for the target task. Further, $\mathcal{S}$ can be used as training data in a fully-supervised setup, either to fine-tune an LLM or to train a conventional model from scratch. Although the dataset $\mathcal{S}$ can be made arbitrarily large, it can quickly hit di-

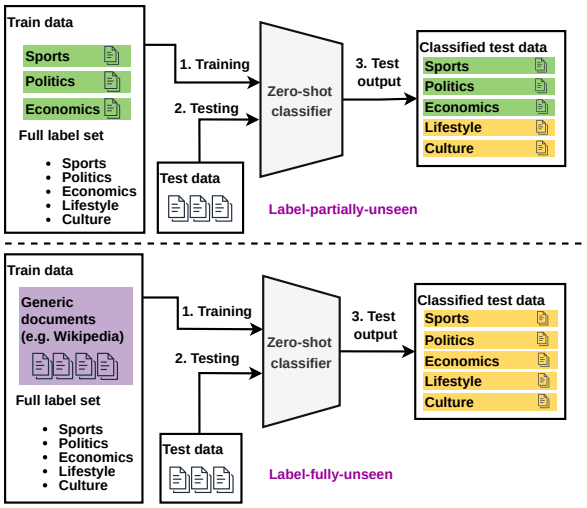

Figure 2: Zero-shot classification variants, as described by Yin et al. (2019). In *label-partially-unseen*, the model is trained on labeled data from a subset of classes, then tested on the full set. In *label-fully-unseen*, the model is trained on unlabeled data, possibly unrelated to the target task. Best viewed in color.

minishing returns, as it tends to contain redundant data (Ye et al., 2022). Moreover, the performance obtained by using $\mathcal{S}$ is not as high as it would be when using a real human-labeled dataset of comparable size. To this end, several efforts have been made to improve the quality of $\mathcal{S}$ (Meng et al., 2022; Ye et al., 2022). A generic dataset, such as Wikipedia, is sometimes employed as a source of additional diversity. Meng et al. (2022) use controlled text generation (Hu et al., 2017) to direct an LLM to generate texts that are relevant for a specific class label. To increase sample diversity, repeating sequences are penalized. Alternatively, Yu et al. (2023) use Wikipedia as a general purpose corpus. In their method, called ReGen, a retrieval model is trained using contrastive learning for the task of finding the most relevant documents from the corpus. ProGen (Ye et al., 2022) and ZeroGen (Ye et al., 2022) represent other alternatives to generate synthetic datasets. ZeroGen uses prompt engineering to generate $\mathcal{S}$ from scratch, without the need of additional data. ProGen employs a feedback loop to iteratively refine a task-specific model.

Some studies relax the zero-shot learning setup in various ways, such as reducing the zero-shot setup to a subset of classes, while providing labeled examples for the others, or considering *the availability of an unlabeled set of samples* from the target classes, but without knowing the corresponding class labels (see Figure 2). For example, Meng et al. (2020) train a model on a subset originating from the same distribution as the dataset used for evaluation, but without labels. This can be a valid setup, depending on the specific problem at hand, mainly when data is available, but labeling is difficult. Gera et al. (2022) adopt a similar approach, based on self-training and entailment, for an instruction model. A model h_0 assigns pseudo-labels based on the confidence of entailment with "This text is about <class>". The resulting candidates are filtered and used to train a new model h_1 that is further used to assign new pseudo-labels, and so on. Although this method may look similar with the one proposed by Yin et al. (2019), Gera et al. (2022) use the entailment task to assign pseudo-labels to a dataset used for downstream training of a target model, while Yin et al. (2019) directly employ entailment for classification.

6 Open-Set Text Classification and Discovery

Trends and approaches. A simple way to address the open-set learning problem is to classify the data using a closed-set algorithm, then estimate how well each sample corresponds to the assigned class. If the confidence (or similarity) is under a certain threshold, the example is considered to belong to the open space. In practice, this threshold can be difficult to set, because, in general, the ratio of instances belonging to C^{uu} in the test set $\mathcal{T}$ cannot be estimated without prior knowledge. This problem is already identified and discussed by Scheirer et al. (2013), who frame the issue as an effort to harmonize the empirical risk, measured on the training data with the unmeasurable *open-space risk*, which is unknown.

There are a few methods that avoid estimating this threshold. Starting from the observation that there are spaces where feature vectors of instances from the same class generally reside close together, some researchers use methods specific to clustering. In this context, an outlierness factor can be calculated to decide if an instance should be classified into a class or allocated to the open space. If an example is an outlier, it probably belongs to the open space. Various outlierness factors have been proposed so far. Walkowiak et al. (2019a) employ the Local Outlier Factor (LOF), which uses a weighted Euclidean distance to find outliers based on the distance to their local neighbors. Subsequent studies of the same authors (Walkowiak et al., 2020, 2019b) use LOF as well. Another factor, called

Angle-Based Outlier Factor, which was originally introduced by Kriegel et al. (2008), is later adapted for text classification by Walkowiak et al. (2018), under the name ABOF2.

Another solution to avoid setting an arbitrary threshold is to identify the examples belonging to the open space before performing the classification task. The problem is therefore broken into two subtasks, outlier detection (OD) and closed-set classification, which can be treated independently. In this setup, the classification part is closed, hence OD is the difficult part.

Open-set solutions via outlier / novelty detection. Kannan et al. (2017) divide the outlier detection algorithms for text into three subcategories: distance-based, density-based, and subspace-based. More recently, energy-based models have also been explored as a promising method for detecting out-of-distribution data. Grathwohl et al. (2019) propose a Joint Energy-Based Model (JEM) by reinterpreting a softmax-based classification network as a generative model. This approach uses the logits to define an energy-based representation of the joint distribution of data points and labels. Another notable contribution to outlier detection is the typicality test, introduced by Nalisnick et al. (2019). This method determines whether a given data point belongs to the model's typical set, rather than relying solely on likelihood estimation. The test is based on the observation that deep generative models can assign higher likelihoods to OOD data than to in-distribution samples. By focusing on typicality instead of raw probability density, this method offers a more robust approach to detecting outliers.

We emphasize that there are two main types of distribution shifts (Baran et al., 2023): *semantic shift*, defined as the occurrence of new classes, and *background shift*, defined as changes that are class-agnostic, such as the level of formality, stance, and so on. If the OD detects background shift, the examples belonging to C^{kk} that exhibit background shift will not even make it to the classifier. Consequently, the outlier detector should only detect semantic shift, but this is difficult without access to outliers during training. The subject of semantic vs. background shift in text is discussed in detail by Arora et al. (2021). While outlier detection in general is very well studied (Wang et al., 2019), there are very few articles focused on text, especially ones trying to isolate semantic shift. Yet, a simple way to detect semantic shift is to use rare terms (Mohotti and Nayak, 2020).

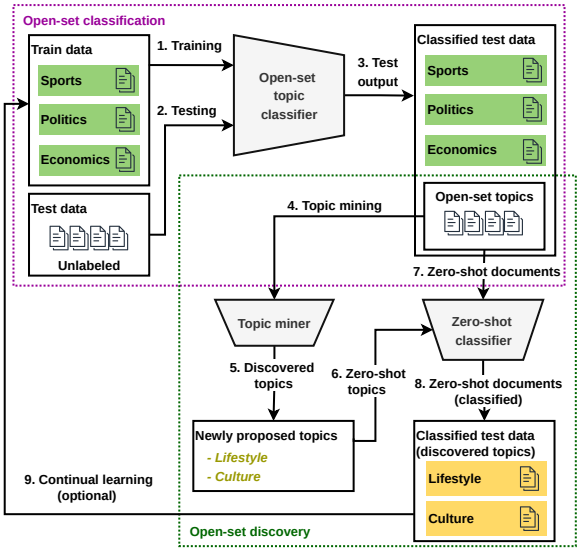

Figure 3: A generic open-set classification and discovery pipeline. The data is initially classified by the open-set classifier. The open-space instances are mined for new class candidates. Then, a zero-shot classifier is used to find the relevant instances. The process can be reiterated via continual learning, including the newly found classes and corresponding documents to retrain the original open-set model. Best viewed in color.

Open-set semi-supervised text classification. Semi-supervised text classification (STC) has its own open-set variant, called open-set semi-supervised text classification (OSTC), introduced by Chen et al. (2023). For the STC task, the training set consists of some labeled examples from each class, and many additional unlabeled examples. In this setup, all the unlabeled examples are presumed to belong to C^{kk}. This assumption is difficult to enforce in practice. In OSTC, the assumption is relaxed such that the unlabeled examples can belong to both C^{kk} and C^{uu}, making the problem open. Since samples from C^{uu} are available during training, we consider that OSTC methods have a clear advantage over open-set learning methods, even though the available samples are unlabeled.

Chen et al. (2024) use adversarial disagreement maximization to increase the difference between in-distribution and OOD examples, improving on their previous solution (Chen et al., 2023). Kim et al. (2022) propose a harder variant of OSTC, that is zero-shot. In this new setup, there are no labeled examples at all. The training data consists of a set of n unlabeled texts $x_1, ..., x_n$ that must be classified into k classes $c_1, ..., c_k$. Each class c_i is only specified through a set of words. The proposed method obtains better results than the corresponding closed-set solutions (Meng et al., 2020; Wang

et al., 2021), when tested in an open-set setting. LLMs have not been extensively studied for OSTC. Chen et al. (2024) are among the few to compare small vs. large language models on OOD detection and OSTC. They found that LLaMA2-7B performs poorly on both tasks, in both zero-shot and few-shot scenarios. Their study shows that prompting LLMs is not sufficient to accurately address OOD detection and OSTC, since a customized approach based on a smaller model can easily surpass zero-shot and few-shot LLM prompting.

Open-set learning and discovery. Ideally, during inference, an open-set system should be able to recognize patterns in the open data and dynamically create new relevant classes that would complement the set of classes C^{kk} provided at train time. A solution to this problem should involve: (i) an *open-set text classification* model to initially split the corpus into C^{kk} and C^{uu}, (ii) a *topic mining* method to identify new relevant topics in the open space, and (iii) a *zero-shot learning* model to classify samples into the newly found topics. A further challenge is posed by the fact that the new classes should be semantically consistent with the existing classes. For example, if the existing classes are "sports", "politics" and "economics", a new class called "culture" is compatible, but one called "breaking news" is not. To obtain semantically equivalent classes, a knowledge base, such as WordNet, might come in handy. A possible architecture for the open-set learning and discovery setup is illustrated in Figure 3. To the best of our knowledge, there is no framework for open-set learning and discovery in the field of text classification.

In the context of open-set text classification and discovery, instruction models can be used to perform the open-set text classification. However, it is difficult to discover new classes via prompting due to the large quantity of text and the relatively short-term memory of LLMs. However, using a reasoning model to interactively refine a list of possible topics is a path worth exploring.

7 Comparison of Learning Frameworks

In Table 2, we compare the three learning frameworks in terms of several key characteristics. From the comparative analysis, we next derive the main limitations of each learning framework.

Learning with background class. The performance depends heavily on the quality and diversity of background samples. Since it is impossible to sample from UUCs, the background data will never fully represent the real open-world. Hence, UUCs will likely be treated as one of the KKCs.

Characteristic	Universum	Zero-shot	Open-set
Can detect OOD data	✓	✗	✓
Set of classes is unbounded	✗	✗	✓
Can classify OOD data	✗	✓	✓
OOD training data required	✓	✗	✗
Open-world assumption	✗	✗	✓
Easy to optimize	✓	✓	✗

Table 2: Comparison of learning frameworks in terms of various characteristics.

Zero-shot learning. Zero-shot models are sensitive to label wording, an aspect that is important to build representative class prototypes for UKCs. Without explicit supervision, the model might struggle to make subtle distinctions between fine-grained UKCs. UUC samples must map to one of the supplied classes, so zero-shot models cannot handle UUCs.

Open-set learning and discovery. The learning problem is hard. It is difficult to define boundaries around KKCs, without examples of UUCs. It is even more difficult to classify UUCs without labeled samples. Moreover, it is hard to solve open-set learning and discovery via a single objective, requiring multiple optimization stages.

8 Conclusion and Future Work

Conclusion. Our survey examined three principal paradigms for text classification under shifting class distributions: learning with background class, zero-shot classification and open-set classification.

Learning with background class relies on auxiliary data to reject out-of-distribution examples. However, its effectiveness is highly sensitive to the quality and distinctiveness of the available KUC samples. When these auxiliary examples fail to capture real-life diversity, the approach struggles to generalize beyond its training set. Zero-shot classification offers flexibility by leveraging semantic representations to extend recognition to unseen classes, yet it is inherently dependent on the robustness of the available semantic knowledge. In practice, when emerging classes deviate significantly from established semantic cues, zero-shot methods can falter, leading to misclassifications or an inability to capture nuanced differences between similar topics. Open-set classification, on the other hand, excels at detecting novelty through effective outlier identification. However, it typically lacks mechanisms for assigning meaningful labels or integrating these outliers into an exist-

ing classification framework. These methodologies have advanced our understanding of handling distribution shifts in text classification, but the field still lacks a unified framework capable of simultaneously discovering, categorizing, and adapting to emerging classes. A promising direction lies in continual learning approaches that can integrate the strengths of all three paradigms, while mitigating their individual weaknesses.

Future directions. Continual learning is likely one of the most promising paths towards unlocking open-set learning and discovery in the text domain. However, adapting continual learning techniques to mitigate open-set learning and discovery requires addressing additional complexities, such as novelty detection, adaptive knowledge retention, and dynamic model scalability. Liu et al. (2023) identify several challenges in the development of an open-word continuous learning system, such as autonomous novelty detection, automatic acquisition of labeled data via interaction, and risk assessment during self-initiated adaptation. One promising direction is disentanglement-based regularization, which separates learned representations into two spaces: a stable and generic one, and a flexible and task-specific one. This method enhances knowledge retention without restricting the adaptability of the model to new tasks, as demonstrated by Huang et al. (2021). Another important avenue is optimizing memory selection for replay, where approaches such as k-means clustering can be used to retain only the most representative examples from previous tasks.

Research on continual learning for natural language generation (NLG) has also provided insights into how adaptive architectures can support evolving textual domains. Yang et al. (2022) propose a transformer calibration mechanism to minimize interference between tasks in continual learning scenarios for NLG. This technique leverages attention calibration and feature recalibration to enable language models to incrementally adapt to new tasks, while preserving previously learned knowledge. Similar strategies could be beneficial for open-set continual learning in text classification, where models must adjust their representation space dynamically as new topics emerge.

Sun et al. (2020) introduce LAMOL (Language Modeling for Lifelong Language Learning), which reformulates various NLP tasks into language modeling. Under this framework, the model generates the replay samples from previously learned tasks, reducing catastrophic forgetting (Kirkpatrick et al., 2017) without storing large historical datasets. Although originally designed for a closed set of tasks, LAMOL could be extended to open-set scenarios by incorporating a novelty detection module. Whenever the system encounters anomalous samples or unfamiliar topics, it can trigger an incremental update process that treats the new categories or subject area as separate tasks. This setup would allow the model to enlarge its internal knowledge base, maintaining prior knowledge, while adapting dynamically to unseen class distributions.

Drawing inspiration from CoLeCLIP (Li et al., 2025b), which introduces open-domain continual learning in the image domain through a combination of task-specific prompting and joint vocabulary learning, we can apply similar principles for text classification in open-set scenarios. The authors propose the development of shared representations and an adaptive prompting mechanism, which jointly enable a model to adjust to new tasks and classes as they emerge, while preserving previously acquired knowledge. Applied to text, this approach translates into designing dynamic prompts and an extensible vocabulary that guide classification under shifting class distributions, simultaneously detecting the emergence of new or unknown topics. For instance, building on the techniques employed by CoLeCLIP, a language model could learn task-specific prompts for different topics, continually adapting and expanding its internal vocabulary as it encounters new terms or concepts. Such an approach would likely need to involve a regularization mechanism to prevent catastrophic forgetting, employing selective memory strategies to retain critical information from past data.

While current open-set methods exhibit difficulties in threshold selection for open-space delimitation, LLMs offer significant potential for addressing class distribution shift. For instance, LLMs can provide auxiliary confidence signals by generating natural language explanations for classification decisions. When an LLM shows high uncertainty or produces inconsistent explanations for a sample, this signals potential inclusion into an unknown class, complementing traditional outlier detection methods. LLMs can also generate semantically coherent labels for discovered clusters (Figure 3, step 4), ensuring consistency with existing classes, e.g. proposing “sports” rather than “tennis” when existing classes are “culture”, “politics” and “economy”.

Acknowledgments

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0090, within PNCDI IV. This research is also supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

Limitations

To the best of our knowledge, this work covers the primary directions pertinent to the task of open-set text classification. However, some relevant studies may have been inadvertently missed due to limited visibility or other constraints.

References

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2022. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20111–20121.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10687–10701.

Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. 2023. Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of the 61st Annual Meeting of the Association For Computational Linguistics Student Research Workshop (SRW)*, pages 119–129.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. 2021. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15313–15322.

Akhil Chaudhary, Evangelos Milios, and Enayat Rajabi. 2024. Top2Label: Explainable zero shot topic labelling using knowledge graphs. *Expert Systems with Applications*, 242:122676.

Junfan Chen, Richong Zhang, Junchi Chen, and Chunming Hu. 2024. Open-set semi-supervised text classification via adversarial disagreement maximization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2170–2180.

Junfan Chen, Richong Zhang, Junchi Chen, Chunming Hu, and Yongyi Mao. 2023. Open-set semi-supervised text classification with latent outlier softening. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–236.

Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 703–711.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*, page 209–218.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 503–509.

Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. 2020. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1021–1030.

Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. 2018. Reducing network agnostophobia. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, pages 9175–9186.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 506–514.

Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. 2021. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1107–1119.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron

Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2019. Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schütze. 2020. Neural topic modeling with continual lifelong learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3907–3917.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2019. Deep anomaly detection with outlier exposure. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yibo Hu and Latifur Khan. 2021. Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 628–636.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1587–1596.

Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. 2021. Continual learning for text classification with information disentanglement based regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2736–2746.

Ramakrishnan Kannan, Hyenkyun Woo, Charu C. Aggarwal, and Haesun Park. 2017. Outlier detection for text data. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, pages 489–497.

Dohyung Kim, Jahwan Koo, and Ung-Mo Kim. 2022. OSP-Class: Open Set Pseudo-labeling with Noise Robust Training for Text Classification. In *Proceedings of the IEEE International Conference on Big Data (BigData)*, pages 5520–5529.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Shu Kong and Deva Ramanan. 2021. OpenGAN: Open-Set Recognition via Open Data Generation . In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 793–802.

Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. 2008. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 444–452.

Ken Lang. 1995. NewsWeeder: Learning to Filter Netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and Christian Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195.

David Lewis. 1987. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository.

Yujie Li, Guannan Lai, Xin Yang, Yonghao Li, Marcello Bonsangue, and Tianrui Li. 2025a. Exploring Open-world Continual Learning with Knowns-Unknowns Knowledge Transfer. *arXiv preprint arXiv:2502.20124*.

Yukun Li, Guansong Pang, Wei Suo, Chenchen Jing, Yuling Xi, Lingqiao Liu, Hao Chen, Guoqiang Liang, and Peng Wang. 2025b. CoLeCLIP: Open-Domain Continual Learning via Joint Task Prompt and Vocabulary Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):15137–15151.

Bing Liu, Sahisnu Mazumder, Eric Robertson, and Scott Grigsby. 2023. AI autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine*, 44(2):185–199.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 38–55.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21464–21475.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–150.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 462–477.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

Wathsala Anupama Mohotti and Richi Nayak. 2020. Efficient outlier detection in text corpus using rare frequency and ranking. *ACM Transactions on Knowledge Discovery from Data*, 14(6):1–30.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Ravi Teja Mullapudi, Fait Poms, William R. Mark, Deva Ramanan, and Kayvon Fatahalian. 2021. Background Splitting: Finding Rare Classes in a Sea of Background. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8039–8048.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. 2019. Detecting out-of-distribution inputs to deep generative models using typicality. In *Proceedings of 4th workshop on Bayesian Deep Learning*.

Hugo Oliveira, Caio Silva, Gabriel L. S. Machado, Keiller Nogueira, and Jefersson A. dos Santos. 2021. Fully convolutional open set segmentation. *Machine Learning*, 112(5):1733–1784.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. 2023. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37.

Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and Q. M. Jonathan Wu. 2023. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4051–4070.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. 2022. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Walter J. Scheirer, Anderson de Rezende Rocha, Arun Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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 (EMNLP)*, pages 1631–1642.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. LAMOL: LAnguage MOdeling for Lifelong Language Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12471–12491.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., Berlin, Heidelberg.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2021. Open-set recognition: A good closed-set classifier is all you need. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. 2012. Classification of short texts by deploying topical annotations. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR)*, pages 376–387.

Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski. 2018. Algorithm based on modified angle-based outlier factor for open-set classification of text documents. *Applied Stochastic Models in Business and Industry*, 34(5):718–729.

Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski. 2019a. Distance Metrics in Open-Set Classification of Text Documents by Local Outlier Factor and Doc2Vec. In *Proceedings of the 32nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)*, pages 102–109.

Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski. 2019b. Open Set Subject Classification of Text Documents in Polish by Doc-to-Vec and Local Outlier Factor. In *Proceedings of the International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, volume 11509, pages 455–463.

Tomasz Walkowiak, Szymon Datko, and Henryk Maciejewski. 2020. Utilizing local outlier factor for open-set classification in high-dimensional data–case study applied for text documents. In *Proceedings of the 2019 Intelligent Systems Conference (IntelliSys)*, pages 408–418.

Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. 2019. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3043–3053.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jason Weston, Ronan Collobert, Fabian Sinz, Léon Bottou, and Vladimir Vapnik. 2006. Inference with the Universum. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 1009–1016.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. 2024. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18297–18307.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. ZeroPrompt: Scaling Prompt-Based Pretraining to 1,000 Tasks Improves Zero-Shot Generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4235–4252.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662.

Peng Yang, Dingcheng Li, and Ping Li. 2022. Continual learning for natural language generations with transformer calibration. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 40–49.

Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. 2025. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5):1–38.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11653–11669.

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 (EMNLP)*, pages 3914–3923.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023. ReGen: Zero-Shot Text Classification via Training Data Generation with Progressive Dense Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey. *arXiv preprint arXiv:2308.10792*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*, pages 649–657.

Jiyang Zheng, Weihao Li, Jie Hong, Lars Petersson, and Nick Barnes. 2022. Towards open-set object detection and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3960–3969.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2024. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*.

Hanzhang Zhou, Zijian Feng, and Kezhi Mao. 2023. Closed boundary learning for classification tasks with

the universum class. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15522–15536.

Fei Zhu, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. 2024. Open-world machine learning: A review and new outlooks. *arXiv preprint arXiv:2403.01759*.

A Appendix

In the supplementary, we discuss related work on continual learning from text and how this research direction is useful in text classification under class distribution shift (Section A.1). We also review the datasets that are commonly used in zero-shot and open-set learning from text (Section A.2).

A.1 Continual Learning from Text

Formal definition. Continual learning (CL) is a paradigm in machine learning focused on enabling systems to continuously learn from a stream of data over time. Unlike traditional approaches that assume a static data distribution and rely on fixed datasets, CL addresses challenges such as retaining knowledge from previous tasks (avoiding catastrophic forgetting (Kirkpatrick et al., 2017)) and adapting to new tasks without extensive retraining (Parisi et al., 2019). CL can be defined as the process of training a model on a sequence of tasks, where each task t can be characterized by a distinct data distribution. At any point in time, the model receives a batch of training samples $D_{t,b} = \{(x, y) \mid x \in \mathcal{X}_{t,b}, y \in \mathcal{Y}_{t,b}\}$, where $\mathcal{X}_{t,b}$ represents a subset of input samples and $\mathcal{Y}_{t,b}$ are the corresponding (task-specific) labels. The task identity is given by $t \in \{1, ..., \infty\}$, while the batch index is represented by $b \in \{1, ..., \infty\}$. Theoretically, the continual learning process continues indefinitely during operation, hence the indexes t and b can grow endlessly. A task t is formally defined by its data distribution $P(\mathcal{X}_t, \mathcal{Y}_t)$ (Wang et al., 2024).

Trends and approaches. CL is a relevant framework for open-set learning, as it can address challenges related to the open-set scenarios, especially in tasks such as text classification by topic when the topic distribution changes. These challenges include handling new classes and adapting to an evolving data distribution. CL can help to discover and learn new topics during inference. Both Chen and Liu (2014) and Gupta et al. (2020) propose frameworks that enable models to accumulate knowledge over time, leveraging past knowledge to improve topic extraction, while mitigating catastrophic forgetting. These approaches integrate lifelong knowledge retention and transfer, ensuring that previously learned topics inform future learning tasks. By incorporating prior knowledge, these models enhance topic coherence, particularly in sparse data scenarios, where limited labeled data can affect the quality of extracted topics. Chen and Liu (2014) introduce the Lifelong Topic Model (LTM), which mines prior knowledge from past document collections and integrates it into new topic modeling tasks through probabilistic inference. Similarly, Gupta et al. (2020) propose the Lifelong Neural Topic Modeling (LNTM) framework, which relies on neural networks and selective data augmentation to balance stability and plasticity. LTM is designed for multi-domain topic extraction, identifying topic structures across different domains, while LNTM is more suitable for continuous document streams, maintaining topic stability, while adapting to novel information. Moreover, LTM employs a knowledge-based topic modeling approach that refines topic quality through extracted knowledge sets, whereas LNTM integrates topic regularization and selective data augmentation to enhance adaptation, while preventing excessive forgetting.

Open-world continual learning. Open-world continual learning is a learning paradigm in which a model incrementally acquires new tasks and, at the same time, detects and integrates previously unseen classes. In their recent work, Li et al. (2025a) propose a framework called HoliTrans, which transfers knowledge from both known and unknown class samples. HoliTrans uses Nonlinear Random Projection to create clearer representations that help to separate new samples from previously learned classes. The method also relies on Distribution-Aware Prototypes, which dynamically adapt as new classes appear. These components help to better distinguish between new and previously seen classes, particularly when unknown classes repeatedly occur.

A.2 Text Datasets for Class Distribution Shift

In Table 3, we provide a list of text classification datasets that are often used in text classification under class distribution shift. These datasets are typically configured for a standard supervised classification setup. We further explain how the listed datasets are modified to serve the zero-shot, open-set, and continual learning setups, respectively.

Dataset	Task	#samples	#classes	Zero-shot	Open-set	Continual
AG News (Zhang et al., 2015)	Topic categorization	127k	4	✓	✓	✓
DBPedia (Lehmann et al., 2014)	Topic categorization	630k	14	✓	✓	✓
20 Newsgroup (Lang, 1995)	Topic categorization	18k	20	✓	✓	✗
Yahoo! Answers (Zhang et al., 2015)	Topic categorization	1.4m	10	✓	✓	✗
SST-2 (Socher et al., 2013)	Polarity classification	69k	2	✓	✗	✗
IMDb (Maas et al., 2011)	Polarity classification	50k	2	✓	✗	✗
Amazon Product Reviews (Chen et al., 2013)	Topic modeling	50k	50	✗	✓	✓
TMNtitle (Gupta et al., 2020)	Topic modeling	32.6k	7	✗	✗	✓
R21578title (Gupta et al., 2020)	Topic modeling	10.8k	90	✗	✗	✓

Table 3: List of datasets that are commonly used in text classification under class distribution shift. For the selection of datasets, we report some of their basic statistics, as well as the tasks where they are commonly employed.

Text datasets for zero-shot classification. A possible methodology for zero-shot testing is proposed by Yin et al. (2019), alongside a few suitable datasets. For topic categorization, the authors propose the Yahoo! Answers (Zhang et al., 2015) dataset. This dataset is reorganized into *dev* and *test* splits, both containing all 10 labels. The *dev* split consists of 6k instances per label, while the *test* split has 10k instances per label. For the labels-partially-unseen scenario, two variants of the dataset are created, with different partitions of seen/unseen classes, both balanced. For labels-fully-unseen there is no training set. Sanh et al. (2022) employ AG News (Zhang et al., 2015) and DBPedia (Lehmann et al., 2014) for zero-shot topic categorization, and IMDb (Maas et al., 2011) for polarity classification. There is no training set for the target task. Similarly, Meng et al. (2020) use the AG News, DBPedia, IMDb and Amazon Product Reviews (Chen et al., 2013) datasets, with all the labels removed. In general, we observe that the preferred datasets are large scale.

Text datasets for open-set classification. The common way to test an open-set solution is to reorganize a dataset designed for supervised text classification. To this end, some classes are reserved to play the role of the open space and are not used during training. While this scenario may seem similar to the zero-shot labels-partially-unseen setup, the labels for the reserved classes are not used. During testing, the system must simply detect them as belonging to the open space, without classifying them into classes. Many works perform multiple experiments, varying the KKCs-to-UUCs ratio between 20% and 100%. It is worth noting that a percentage of 100% known classes is not a closed-set problem, because the model can still classify instances as open space.

Kim et al. (2022) report experiments on AG News and DBPedia. The KKCs-to-UUCs ratio is chosen among three values: 25%, 50% and 75%. For AG News, all the combinations of known / unknown classes having at least one known and one unknown class are compared. Some studies (Chen et al., 2023, 2024) conduct experiments on three datasets, namely AG News, DBPedia and Yahoo! Answers, reserving two, four or six classes to play the role of UUCs. Other papers (Walkowiak et al., 2018, 2019b) use custom datasets, specifically tailored for the open-set task. Walkowiak et al. (2019b) are the only ones who use data in a foreign language, namely Polish.

Text datasets for continual learning. Gupta et al. (2020) report experiments on Amazon Product Reviews (Chen et al., 2013), 20NSshort, TMNtitle and R21578title datasets. They use these datasets as future tasks in a continual learning framework for topic modeling, testing the ability of models to learn from short texts, retain knowledge, and adapt to new domains without catastrophic forgetting. The Amazon Product Reviews dataset is preprocessed according to the methodology described by Chen et al. (2013). The preprocessing includes sentence detection, lemmatization, and POS tagging, as well as the removal of punctuation, stop words, and rare terms. To avoid excessive thematic overlap, the domain name of each collection is removed. 20NSshort is a subset of 20 Newsgroups (Lang,

1995), containing only documents with fewer than 20 words, making topic inference challenging due to sparse content. Similarly, R21578title is derived from Reuters-21578 (Lewis, 1987), retaining only article titles, further limiting contextual information. TMNtitle comes from Tag My News (TMN) (Vitale et al., 2012) and includes only news headlines, reflecting real-world short-text classification tasks.

Sun et al. (2020) use DBPedia as part of a sequential continual learning setup alongside other datasets, such as AG News, Amazon Product Reviews, and Yahoo! Answers. As the authors sequentially introduce new classification tasks, DBPedia helps them to assess how well the model could retain previously learned categories, while adapting to new ones.

General remark. We find that datasets used in literature do not always reflect real-world open-set scenarios. This is mostly due to the fact that nearly all datasets were introduced more than 10 years ago. We thus believe that collecting new resources that cover realistic open-set setups is an important avenue for future research.