A Survey on Out-of-Distribution Detection in NLP

Anonymous authors
Paper under double-blind review

Abstract

Out-of-distribution (OOD) detection is essential for the reliable and safe deployment of machine learning systems in the real world. Great progress has been made over the past years. This paper presents the first review of recent advances in OOD detection with a particular focus on natural language processing approaches. First, we provide a formal definition of OOD detection and discuss several related fields. We then categorize recent algorithms into three classes according to the data they used: (1) OOD data available, (2) OOD data unavailable + in-distribution (ID) label available, and (3) OOD data unavailable + ID label unavailable. Third, we introduce datasets, applications, and metrics. Finally, we summarize existing work and present potential future research topics.

1 Introduction

Natural language processing systems deployed in the wild often encounter out-of-distribution (OOD) samples that are not seen in the training phase. For example, a natural language understanding (NLU) component in a functional dialogue system is typically developed using a limited training set that encompasses a finite number of intents. However, when deployed, this NLU component may be exposed to an endless variety of user inputs, some of which (i.e., OOD samples) may include intents not supported by the training. A reliable and trustworthy NLP model should not only obtain high performance on samples from seen distributions, i.e., In-distribution (ID) samples, but also accurately detect OOD samples (Amodei et al., 2016; Boult et al., 2019). For instance, when building task-oriented dialogue systems, it is hard, if not impossible, to cover all possible user intents in the training stage. It is critical for a practical system to detect these OOD intents or classes in the testing phase so that they can be properly handled (Zhan et al., 2021).

However, existing flourishes of neural-based NLP models are built upon the *closed-world assumption*, i.e., the training and testing data are sampled from the same distribution (Vapnik, 1991). This assumption is often violated in practice, where deployed models are generally confronting an *open-world*, i.e., some testing data may come from OOD distributions that are not seen in training (Bendale & Boult, 2015; Fei & Liu, 2016). It is also worth noting that although large language models (LLMs) have exhibited superior performance in various tasks by training on an enormous set of texts, the knowledge exhibited in these training texts are limited to a certain cut-off date. OOD detection is still an important task for these LLMs since the world is involving. New tasks may be developed after the knowledge cut-off date.

A rich line of work has been proposed to tackle problems introduced by OOD samples. Specifically, distributional shifts in NLP can be broadly divided into two types: 1. semantic shift, i.e., OOD samples may come from unknown categories, and therefore should not be blindly predicted into a known category; 2. non-semantic shift, i.e., OOD samples may come from different domains or styles but share the same semantic with some ID samples (Arora et al., 2021). The detection of OOD samples with semantic shift is the primary focus of this survey, where the label set $\mathcal{Y}$ of ID samples is different from that of OOD samples. The ability of detecting OOD samples is critical for building safe NLP systems for, say, text classification (Hendrycks & Gimpel, 2016), question answering (Kamath et al., 2020), and machine translation (Kumar & Sarawagi, 2019).

Although there already exists surveys on many aspects of OOD, such as OOD generalization (Wang et al., 2022) and OOD detection in computer vision (CV) (Yang et al., 2021), *a comprehensive survey for OOD* *detection in NLP is still lacking and thus urgently needed for the field.* Concretely, applying OOD detection to NLP tasks requires specific considerations, e.g., tackling discrete input spaces, handling complex output structures, and considering contextual information, which have not been thoroughly discussed. Our key contributions are summarized as follows:

1. We propose a novel taxonomy of OOD detection methods based on the availability of OOD data (Section 3) and discuss their pros and cons for different settings (Section 6.1).

2. We present a survey on OOD detection in NLP and identify various differences between OOD detection in NLP and CV (Section 6.3).

3. We review datasets, applications (Section 4), metrics (Section 5), and future research directions (Section 6.4) of OOD detection in NLP.

2 OOD Detection and Related Areas

Definition 1 (Data distribution). *Let $\mathcal{X}$ denote a nonempty input (non-semantic) space and $\mathcal{Y}$ a label (semantic) space. A data distribution is defined as a joint distribution $P(X,Y)$ over $\mathcal{X} \times \mathcal{Y}$. $P(X)$ and $P(Y)$ refer to the marginal distributions for inputs and labels, respectively.*

In practice, common non-semantic distribution shifts on $P(X)$ include domain shifts (Wang et al., 2022), sub-population shifts (Koh et al., 2021), style changes (Pavlick & Tetreault, 2016), or adversarial examples (Carlini & Wagner, 2017; Rozsa et al., 2017). Typically, the label space $\mathcal{Y}$ remains unchanged in these non-semantic shifts, and sophisticated methods are developed to improve the model's robustness and generalization performance (Hendrycks et al., 2020). On the contrary, semantic distribution shifts on $P(Y)$ generally lead to a new label space $\widetilde{\mathcal{Y}}$ that are different from the one seen in the training phase (Bendale & Boult, 2016). These shifts are usually caused by the occurrence of new classes at the testing stage. In this work, we mainly focus on detecting OOD samples with semantic shifts, the formal definition of which is given as follows:

Definition 2 (OOD detection). *We are given an ID training set $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L} \sim P_{train}(X,Y)$, where $\mathbf{x}_i \in \mathcal{X}_{train}$ is a training instance, and $y_i \in \mathcal{Y}_{train} = \{1, 2, ..., K\}$ is the associated class label. Facing the emergence of unknown classes, we are given a test set $\mathcal{D}_{test} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \sim P_{test}(X,Y)$, where $\mathbf{x}_i \in \mathcal{X}_{test}$, and $y_i \in \mathcal{Y}_{test} = \{1, ..., K, K+1\}$. Note that class $K+1$ is a group of novel categories representative of OOD samples, which may contain more than one class. The overall goal of OOD detection is to learn a predictive function f from $\mathcal{D}_{train}$ to achieve a minimum expected risk on $\mathcal{D}_{test}$: $\min_f \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{test}} \mathbb{I}(y \neq f(\mathbf{x}))$, i.e., not only classify known classes but also detect the unknown categories.*

We briefly describe the related research areas:

Domain generalization (DG) (Wang et al., 2022), or out-of-distribution generalization, aims to learn a model from one or several source domains and expect these learned models to generalize well on unseen testing domains (i.e., target domains). DG mainly focuses on the non-semantic drift, i.e., the training and testing tasks share the same label space $\mathcal{Y}$ while they have different distributions over the input space $\mathcal{X}$. Different from DG, OOD detection handles a different label space during testing.

Domain adaptation (DA) (Blitzer et al., 2006) follows most settings of DG except that DA has access to some unlabeled data from the target domain in the training process (Ramponi & Plank, 2020). Similar to DG, DA also assumes the label space remains unchanged.

Zero-shot learning (Wang et al., 2019) aims to use learned models to classify samples from unseen classes. The main focus of zero-shot learning is to obtain the correct labels for these unseen classes. However, OOD detection in general only needs to detect samples from unseen classes without further classifying them. Some OOD detection models can also classify samples from seen classes since these samples are annotated in the training set.

Meta-learning (Vilalta & Drissi, 2002) aims to learn from the model training process so that models can quickly adapt to new data. Different from meta-learning, achieving strong few-shot performance is not the

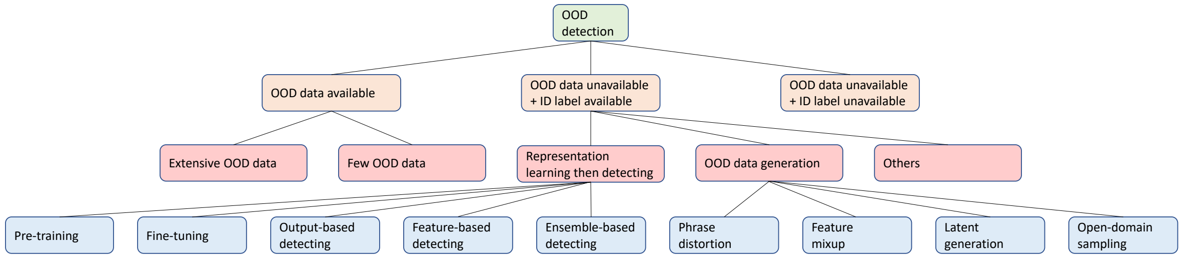

Figure 1: Taxonomy of OOD detection methods.

major focus of OOD detection. Nevertheless, the idea of meta-learning can serve as a strategy for OOD detection (Xu et al., 2019; Li et al., 2021) by simulating the behaviors of predicting unseen classes in the training stage.

Positive-unlabeled Learning (Zhang & Zuo, 2008), or PU learning, aims to train a classifier with only positive and unlabeled examples while being able to distinguish both positive and negative samples in testing. However, OOD detection considers multiple classes in training. PU learning approaches can be applied to tackle the OOD detection problem when only one labeled class exists (Li & Liu, 2003).

Transfer Learning (Ruder et al., 2019) aims to leverage data from additional domains or tasks to train a model with better generalization properties. Most transfer learning approaches target at producing robust representations that are agnostic of their downstream tasks. OOD detection can be regarded as a downstream task for transfer learning.

3 Methodology

A major challenge of OOD detection is the lack of representative OOD data, which is important for estimating OOD distributions (Zhou et al., 2021b). As shown in Figure 1, we classify existing OOD detection methods into three categories according to the availability of OOD data. Methods covered in our survey are selected following the criteria listed in Appendix A.

3.1 OOD Data Available

Methods in this category assume access to both labeled ID and OOD data during training. Based on the quantity and diversity of OOD data, we further classify these methods into two subcategories:

3.1.1 Detection with Extensive OOD Data

Some methods assume that we can access extensive OOD data in the training process together with ID data. In this subcategory, one line of work formulates OOD detection as a discriminative classification task, i.e., a special label is allocated in the label space for OOD samples. Fei & Liu (2016); Larson et al. (2019) formed a $(K+1)$-way classification problem, where K denoted the number of ID classes and the $(K+1)^{th}$ class represented OOD samples. Larson et al. (2019); Kamath et al. (2020) regarded OOD detection as a binary classification problem, where the two classes correspond to ID and OOD samples, respectively. Kim & Kim (2018) introduced a neural joint learning model with a multi-class classifier for domain classification and a binary classifier for OOD detection.

Another line of work optimizes an outlier exposure regularization term on these OOD samples to refine the representations and OOD scores learned by the OOD detector. Hendrycks et al. (2018) introduced a generalized outlier exposure (OE) loss to train models on both ID and OOD data. For example, when using the maximum Softmax probability detector (Hendrycks & Gimpel, 2016), the OE loss pushes the predicted distribution of OOD samples to a uniform distribution (Lee et al., 2018a). When the labels of ID data are not available, the OE loss degenerates to a margin ranking loss on the predicted distributions of ID and OOD samples. Zeng et al. (2021b) added an entropy regularization objective to enforce the predicted distributions of OOD samples to have high entropy.

3.1.2 Detection with Few OOD Data

Some methods assume that we can only access a small amount of OOD data besides ID data. This setting is more realistic in practice since it is expensive to annotate large-scale OOD data. Several methods in this subcategory are developed to generate pseudo samples based on a small number of seed OOD data. Chen & Yu (2021) constructed pseudo-labeled OOD candidates using samples from an auxiliary dataset and kept only the most beneficial candidates for training through a novel election-based filtering mechanism. Rather than directly creating OOD samples in natural language, Zeng et al. (2021b) borrowed the idea of adversarial attack (Goodfellow et al., 2014) to obtain model-agnostic worst-case perturbations in the latent space, where these perturbations or noise can be regarded as augmentations for OOD samples. Note that techniques used by these methods with few OOD data (i.e., increasing the diversity and quantity of OOD data) may also help the detection methods with extensive OOD data (Shu et al., 2021). See Section 4 for the construction schemes of OOD samples in the testing stage and Appendix B for more details of common OOD detection datasets.

3.2 OOD Data Unavailable + ID Label Available

Building OOD detectors using only labeled ID data is the major focus of research communities. We generally classify existing literature into three subcategories based on their learning principles:

3.2.1 Learn Representations Then Detect

Some methods formulize the OOD detector f into two components: a representation extractor g and an OOD scoring function d, i.e., $f(\mathbf{x}) = d(g(\mathbf{x}))$: g aims to capture a representation space $\mathcal{H}$ in which ID and OOD samples are distinct, and d maps each extracted representation into an OOD score so that OOD samples can be detected based on a selected threshold. These two approaches are generally perpendicular to each other. We provide an overview of methods to enhance these two components:

a. Representation Learning usually involves two stages: (1) a *pre-training* stage leverages massive unlabeled text corpora to extract representations that are suitable for general NLP tasks; (2) a *fine-tuning* stage uses labeled in-domain data to refine representations for specified downstream tasks. An overview of these two stages is given here:

Pre-training Pre-trained transformer models such as BERT (Kenton & Toutanova, 2019) have become the de facto standard to implement text representation extractors. Hendrycks et al. (2020) systematically measured the OOD detection performance on various representation extractors, including bag-of-words models, ConvNets (Gu et al., 2018), LSTMs (Hochreiter & Schmidhuber, 1997), and pre-trained transformer models (Vaswani et al., 2017). Their results show that pre-trained models achieve the best OOD detection performance, while the performances of all other models are often worse than chance. The success of pre-trained models attributes to these diverse corpora and effective self-supervised training losses used in training (Hendrycks et al., 2019).

Moreover, it is observed that better-calibrated models generally produce higher OOD detection performance Lee et al. (2018a). Desai & Durrett (2020) evaluated the calibration of two pre-trained models, BERT and RoBERTa (Liu et al., 2019), on different tasks. They found that pre-trained models were better calibrated in out-of-domain settings, where non-pre-trained models like ESIM (Chen et al., 2017) were overconfident. Dan & Roth (2021) also demonstrated that larger pre-trained models are more likely to be better calibrated and thus result in higher OOD detection performance.

Fine-tuning With the help of labeled ID data, various approaches are developed to fine-tune the representation extractor to widen margins between ID and OOD samples. Lin & Xu (2019) proposed a large margin cosine loss (LMCL) to maximize the decision margin in the latent space. LMCL simultaneously maximizes inter-class variances and minimizes intra-class variances. Yan et al. (2020) introduced a semantic-enhanced

Gaussian mixture model to enforce ball-like dense clusters in the feature space, which injects semantic information of class labels into the Gaussian mixture distribution.

Zeng et al. (2021a); Zhou et al. (2021b) proposed a contrastive learning framework (Chen et al., 2020) to increase the discrepancy for representations extracted from different classes. They hypothesized that increasing inter-class discrepancies helps the model learn discriminative features for ID and OOD samples and therefore improves OOD detection performances. Concretely, a supervised contrastive loss (Khosla et al., 2020; Gunel et al., 2020) and a margin-based contrastive loss was investigated. Zeng et al. (2021b) proposed a self-supervised contrastive learning framework to extract discriminative representations of OOD and ID samples from unlabeled data. In this framework, positive pairs are constructed using the back-translation scheme. Zhou et al. (2022) applied KNN-based contrastive learning losses to OOD detectors and Wu et al. (2022) used a reassigned contrastive learning scheme to alleviate the over-confidence issue in OOD detection.

Moreover, there are some regularized fine-tuning schemes to tackle the over-confidence issue of neural-based OOD detectors. Kong et al. (2020) addressed this issue by introducing an off-manifold regularization term to encourage producing uniform distributions for pseudo off-manifold samples. Shen et al. (2021) designed a novel domain-regularized module that is probabilistically motivated and empirically led to a better generalization in both ID classification and OOD detection.

b. OOD Scoring processes usually involve a scoring function d to map the representations of input samples to OOD detection scores. A higher OOD score indicates that the input sample is more likely to be OOD. The implementation of d can be generally categorized into three types: (1) *output-based detecting*, (2) *feature-based detecting*, and (3) *ensemble-based detecting*:

Output-based Detecting compute the OOD score based on the predicted probabilities. Hendrycks & Gimpel (2016); Hendrycks et al. (2020) used the maximum Softmax probability as the detection score, and Liang et al. (2018) improved this scheme with the temperature scaling approach. Shu et al. (2017) employed K 1-vs-rest Sigmoid classifiers for K predefined ID classes and used the maximum probabilities from these classifiers as the detection score. Liu et al. (2020) proposed an energy score for better distinguishing ID/OOD samples. The energy score is theoretically aligned with the probability density of the inputs.

Feature-based Detecting leverages features derived from intermediate layers of the model to implement density-based and distance-based scoring functions. Gu et al. (2019) proposed a nearest-neighbor based method with a distance-to-measure metric. Breunig et al. (2000) used a local outlier factor as the detection score, in which the concept "local" measured how isolated an object was with respect to surrounding neighborhoods. Lee et al. (2018b); Podolskiy et al. (2021) obtained the class-conditioned Gaussian distributions with respect to features of the deep models under Gaussian discriminant analysis. This scheme resulted in a confidence score based on the Mahalanobis distance. While Mahalanobis imposes a strong distributional assumption on the feature space, Sun et al. (2022) demonstrated the efficacy of non-parametric nearest neighbor distance for OOD detection. Zhang et al. (2021) proposed a post-processing method to learn an adaptive decision boundary (ADB) for each ID class. Specifically, the ADB is learned by balancing both the empirical and open space risks (Scheirer et al., 2014). Recently, Ren et al. (2022) proposed to detect OOD samples for conditional language generation tasks (such as abstractive summarization and translation) by calculating the distance between testing input/output and a corresponding background model in the feature space.

Ensemble-based Detecting uses predictive uncertainty of a collection of supporting models to compute OOD scores. Specifically, an input sample is regarded as an OOD sample if the variance of these models' predictions is high. Gal & Ghahramani (2016) modeled uncertainties by applying dropouts to neural-based models. This scheme approximates Bayesian inference in deep Gaussian processes. Lakshminarayanan et al. (2017) used deep ensembles for uncertainty quantification, where multiple models with the same architecture were trained in parallel with different initializations. Lukovnikov et al. (2021) further proposed a heterogeneous ensemble of models with different architectures to detect compositional OOD samples for semantic parsing.

3.2.2 Generate Pseudo OOD Samples

A scheme to tackle the problem of lacking OOD training samples is to generate pseudo OOD samples during training (Lang et al., 2022). With these generated pseudo OOD samples, OOD detectors can be solved by methods designed for using both labeled ID and OOD data. There are mainly four types of approaches to generate pseudo OOD samples: (1) *phrase distortion*, (2) *feature mixup*, (3) *latent generation*, and (4) *open-domain sampling*:

Phrase Distortion approaches generate pseudo OOD samples for NLP tasks by selectively replacing text phrases in ID samples. Ouyang et al. (2021) proposed a data manipulation framework to generate pseudo OOD utterances with importance weights. Choi et al. (2021) proposed OutFlip, which revised a white-box adversarial attack method HotFlip to generate OOD samples. Shu et al. (2021) created OOD instances from ID examples with the help of a pre-trained language model.

Feature Mixup strategy (Zhang et al., 2018) is also a popular technique for pseudo data generation. Zhan et al. (2021) generated OOD samples by performing linear interpolations between ID samples from different classes in the representation space. Zhou et al. (2021a) leveraged the manifold Mixup scheme (Verma et al., 2019) for pseudo OOD sample generation. Intermediate layer representations of two samples from different classes are mixed using scalar weights sampled from the Beta distribution. These feature-mixup-based methods achieved promising performance while remaining conceptually and computationally straightforward.

Latent Generation approaches considered to use generative adversarial networks (GAN) (Goodfellow et al., 2020) to produce high-quality pseudo OOD samples. Lee et al. (2018a) proposed to generate boundary samples in the low-density area of the ID distribution as pseudo-OOD samples. Ryu et al. (2018) built a GAN on ID data and used the discriminator to generate OOD samples in the continuous feature space. Zheng et al. (2020) generated pseudo OOD samples using an auto-encoder with adversarial training in the discrete text space. Marek et al. (2021) proposed OodGAN, in which a sequential generative adversarial network (SeqGAN) (Yu et al., 2017) was used for OOD sample generation. This model follows the idea of Zheng et al. (2020) but works directly on texts and hence eliminates the need to include an auto-encoder.

Open-domain Sampling approaches directly uses sentences from other corpora as pseudo OOD samples (Zhan et al., 2021).

3.2.3 Other Methods

We also review some representative methods that do not belong to the above two categories. Vyas et al. (2018) proposed to use an ensemble of classifiers to detect OOD, where each classifier was trained in a self-supervised manner by leaving out a random subset of training data as OOD data. Li et al. (2021) proposed kFolden, which included k classifiers for k class labels. Each classifier was trained on a subset with $k-1$ classes while leaving one class unknown. Tan et al. (2019) tackled the problem of OOD detection with limited labeled ID training data and proposed an OOD-resistant Prototypical Network to build the OOD detector. Ren et al. (2019); Gangal et al. (2020) used the likelihood ratio produced by generative models to detect OOD samples. The likelihood ratio effectively corrects confounding background statistics for OOD detection. Ryu et al. (2017) employed the reconstruction error as the detection score.

3.3 OOD data unavailable + ID label unavailable

OOD detection using only unlabeled ID data can be used for non-classification tasks. In fact, when ID labels are unavailable, our problem setting falls back to the classic anomaly detection problem, which is developed with a rich set of literature (Pang et al., 2021; Chalapathy & Chawla, 2019). However, this problem setting is rarely investigated in NLP studies. We keep this category here for the completeness of our survey while leaning most of our focus on NLP-related works.

Methods in this category mainly focus on extracting more robust features and making a more accurate estimation for the data distribution. Zong et al. (2018) proposed a DAGMM model for unsupervised OOD detection, which utilized a deep auto-encoder to generate low-dimensional representations to estimate OOD scores. Xu et al. (2021) transformed the feature extracted from each layer of a pre-trained transformer model into one low-dimension representation based on the Mahalanobis distance, and then optimized an OC-SVM for detection. Some works also use language models (Nourbakhsh & Bang, 2019) and word representations Bertero et al. (2017) to detect OOD inputs on various tasks such as log analysis (Yadav et al., 2020) and data mining Agrawal & Agrawal (2015).

4 Datasets and Applications

In this section, we briefly discuss representative datasets and applications for OOD detection. We classify existing OOD detection datasets into three categories according to the construction schemes of OOD samples in the testing stage:

(1) Annotate OOD Samples: This category of datasets contains OOD samples that are manually annotated by crowd-source workers. Specifically, CLINIC150 (Larson et al., 2019) is a manually labeled single-turn dialogue dataset that consists of 150 ID intent classes and 1,200 out-of-scope queries. STAR (Mosig et al., 2020) is a multi-turn dialogue dataset with annotated turn-level intents, in which OOD samples are labeled as "out_of_scope", "custom", or "ambiguous". ROSTD (Gangal et al., 2020) is constructed by annotating about 4,000 OOD samples on the basis of the dataset constructed by Schuster et al. (2019).

(2) Curate OOD samples using existing classes: This category of datasets curates OOD examples by holding out a subset of classes in a given corpus (Zhang et al., 2021). Any text classification datasets can be adopted in this process.

(3) Curate OOD samples using other corpora: This category of datasets curates OOD samples using samples extracted from other datasets (Hendrycks et al., 2020; Zhou et al., 2021b), i.e., samples from other corpora are regarded as OOD samples. In this way, different NLP corpora can be combined to construct OOD detection tasks.

OOD detection tasks have also been widely applied in various NLP applications. We generally divide these applications into two types:

(1) Classification Tasks are natural applications for OOD detectors. Almost every text classifier built in the closed-world assumption needs the OOD detection ability before deploying to production. Specifically, intent classification for dialogue systems is the most common application for OOD detection (Larson et al., 2019; Lin & Xu, 2019). Other popular application scenarios involve general text classification (Zhou et al., 2021b; Li et al., 2021), sentiment analysis (Shu et al., 2017), and topic prediction (Rawat et al., 2021).

(2) Selective Prediction Tasks predict higher-quality outputs while abstaining on uncertain ones (Geifman & El-Yaniv, 2017; Varshney et al., 2022). This setting can be combined naturally with OOD detection techniques. A few studies use OOD detection approaches for selective prediction in question answering, semantic equivalence judgments, and entailment classification (Kamath et al., 2020; Xin et al., 2021).

5 Metrics

The main purposes of OOD detectors are separating OOD and ID input samples, which is essentially a binary classification process. Most methods mentioned above try to compute an *OOD score* for this problem. Therefore, threshold-free metrics that are generally used to evaluate binary classifiers are commonly used to evaluate OOD detectors:

AUROC: Area Under the Receiver Operating Characteristic curve Davis & Goadrich (2006). The Receiver Operating Characteristic curve is a plot showing the true positive rate $TPR = \frac{TP}{TP+FN}$ and the false positive rate $FPR = \frac{FP}{FP+TN}$ against each other, in which TP, TN, FP, FN denotes true positive, true negative, false positive, false negative, respectively. For OOD detection tasks, ID samples are usually regarded as positive. Specifically, a random OOD detector yields an AUROC score of 50% while a "perfect" OOD detector pushes this score up to 100%.

AUPR: Area Under the Precision-Recall curve Manning & Schutze (1999). The Precision-Recall curve plots the precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$ against each other. The metric AUPR is used when the positive and negative classes in the testing phase are severely imbalanced because the metric AUROC is biased in this situation. Generally, two kinds of AUPR scores are reported: 1) **AUPR-IN** where ID samples are specified as positive; 2) **AUPR-OUT** where OOD samples are specified as positive.

Besides these threshold-free metrics, we are also interested in the performance of OOD detectors after the deployment, i.e., when a specific threshold is selected. The following metric is usually used to measure this performance:

FPR@N: The value of FPR when TPR is $N\%$ Liang et al. (2018); Lee et al. (2018a). This metric measures the probability that an OOD sample is misclassified as ID when the TPR is at least $N\%$. Generally, we set $N = 95$ or $N = 90$ to ensure high performance on ID samples. This metric is important for a deployed OOD detector since obtaining a low FPR score while achieving high ID performance is important for practical systems.

In addition to the ability to detect OOD samples, some OOD detectors are also combined with downstream ID classifiers. Specifically, for a dataset that contains K ID classes, these modules allocate an additional OOD class for all the OOD samples and essentially perform a $K + 1$ class classification task. The following metrics are used to evaluate the overall performance of these modules:

F1: The macro F1 score is used to evaluate classification performance, which keeps the balance between precision and recall. Usually, F1 scores are calculated over all samples to estimate the overall performance. Some studies also compute F1 scores over ID and OOD samples, respectively, to evaluate fine-grained performances (Zhang et al., 2021).

Acc: The accuracy score is also used to evaluate classification performance Zhan et al. (2021). See Appendix C for more details of various metrics.

6 Discussion

6.1 Pros and Cons for Different Settings

Labeled OOD data provide valuable information for OOD distributions, and thus models trained using these OOD samples usually achieve high performance in different applications. However, the collection of labeled OOD samples requires additional efforts that are extremely time-consuming and labor-intensive. Moreover, due to the infinite compositions of language, it is generally impractical to collect OOD samples for all unseen cases. Using only a small subset of OOD samples may lead to serious selection bias issues and thus hurt the generalization of the learned model. Therefore, it is important to develop OOD detection methods that do not rely on labeled OOD samples.

OOD detection using only labeled ID data fits the above requirements.in The representation learning and detecting approaches decompose the OOD detection process in this setting into two stages so that we can separately optimize each stage. Specifically, the representation learning stage attempts to learn distinct feature spaces for ID/OOD samples. Results show that this stage benefits from recent advances in pre-training and semi-supervised learning schemes on unlabeled data. Recent research also shows that a good ID classifier benefits the OOD detection Vaze et al. (2021). OOD scoring functions aim to produce reliable scores for OOD detection. Various approaches generate the OOD score with different distance measurements and distributions. Another way to tackle the problem of lacking annotated OOD data is to generate pseudo OOD samples. Approaches in this category benefit from the strong language modeling prior and the generation ability of pre-trained models. Promising results are reported by applying the mixup strategy.

In some applications, we can only obtain a set of ID data without any labels. This situation is commonly encountered in non-classification tasks where we also need to detect OOD inputs. Compared to NLP, this setting is more widely investigated in other fields like machine learning and computer vision (CV). Popular approaches involve using estimated distribution densities or reconstruction losses as the OOD scores.

6.2 Large Language Models for OOD Detection

Recent progress in large language models (LLMs) has led to quality approaching human-performance on research datasets and thus LLMs dominate the NLP field (Brown et al., 2020; Bommasani et al., 2021). With LLMs, many NLP tasks such as text summarization, semantic parsing, and translation can be formulated as a general "text to text" task and have achieved promising results (Raffel et al., 2020; Zhang et al., 2020). In this setting, OOD samples are assumed to be user inputs that significantly deviate from the training data distribution (Xu et al., 2021; Lukovnikov et al., 2021; Ren et al., 2022). These OOD inputs should also be detected because many machine learning models can make overconfident predictions for OOD inputs, leading to significant AI safety issues (Hendrycks & Gimpel, 2016; Ovadia et al., 2019). Moreover, language models are typically trained to classify the next token in an output sequence and may suffer even worse degradation on OOD inputs as the classification is done auto-regressively over many steps. Hence, it is important to know when to trust the generated output of LLMs (Si et al., 2022).

In parallel, LLMs embed broad-coverage world knowledge that can help a variety of downstream tasks (Petroni et al., 2019). Recently, Dai et al. (2023b) apply world knowledge from LLMs to multi-modal OOD detection (Ming et al., 2022) by generating descriptive features for ID class names (Menon & Vondrick, 2023), which significantly increase the OOD detection performance. Meanwhile, LLMs can be explored for text data augmentation generally (Dai et al., 2023a), which could enhance OOD performance by generating diverse, high-quality OOD training data.

6.3 Comparison between NLP and CV in OOD Detection

OOD detection is an active research field in CV communities (Yang et al., 2021) and comprehensive OOD detection benchmarks in CV are constructed (Yang et al., 2022a). A few OOD detection approaches for NLP tasks are remolded from CV research and thus these approaches share a similar design. However, NLU tasks have different characteristics compared to CV tasks. For example, models in NLP need to tackle discrete input spaces and handle complex output structures. Therefore, additional efforts should be paid to develop algorithms for OOD detection in NLP. Although this paper mainly focuses on NLP tasks, it is beneficial to give more discussion about the OOD detection algorithms designed for NLP and CV tasks. Specifically, we summarize the differences in OOD detection between NLP and CV in the following three aspects:

Discrete Input NLP handles token sequences that lie in discrete spaces. Therefore distorting ID samples in their surface space (Ouyang et al., 2021; Choi et al., 2021; Shu et al., 2021) produces high-quality OOD samples if a careful filtering process is designed. On the contrary, CV tackles inputs from continuous spaces, where it is hard to navigate on the manifold of the data distribution. Du et al. (2022b;a) showed OOD synthesizing in the pixel space with a noise-additive manner led to limited performance.

Complex Output Most OOD detection methods in CV are proposed for K-way classification tasks. However, in NLP, conditional language generation tasks need to predict token sequences that lie in sequentially structured distributions, such as semantic parsing (Lukovnikov et al., 2021), abstractive summarization, and machine translation (Ren et al., 2022). Hence, the perils of OOD are arguably more severe as (a) errors may propagate and magnify in sequentially structured output, and (b) the space of low-quality outputs is greatly increased as arbitrary text sequences can be generated. OOD detection methods for these conditional language generation tasks should consider the internal dependency of input-output samples.

Contextual Information Some datasets in NLP contain contextual information. It is important to properly model this extra information for OOD detection in these tasks. For example, STAR (Mosig et al., 2020) is a multi-turn dialogue dataset, and effective OOD detectors should consider multi-turn contextual knowledge in their modeling process (Chen & Yu, 2021). However, most CV models only consider single images as their inputs.

6.4 Future Research Challenges

OOD Detection and Domain Generalization In most practical applications, we are not only interested in detecting OOD inputs that are semantically shifted, but also required to build more robust ID classifiers that can tackle covariate shifted data Yang et al. (2021). We believe there are opportunities to tackle problems of OOD detection and domain generalization in a unified framework. Recent work in CV also shows that OOD detection and OOD generalization can be optimized in a unified margin-based framework (Bai et al., 2023). Future research opportunities can be explored to equip OOD detectors with better text representation extractors since recent results demonstrate that a good ID classifier improves the OOD detection performance (Vaze et al., 2021). Both new task design and algorithm development can be investigated.

OOD Detection with Extra Information Sources Humans usually consider OOD inputs easily distinguishable because they can access external information besides plain texts (e.g., images, audio, and videos). OOD detectors are expected to perform better if we can equip them with inputs from different sources or modalities. Although various works are proposed to model each single information source, such as text or image, few works are dedicated to combining different sources, and no studies try to equip OOD detectors with external knowledge, such as structured knowledge graphs. We envision great performance improvements if we can properly model external knowledge or multi-modality inputs in OOD detectors. Also note that this research direction still lies in the scope of our taxonomy shown in Figure 1 since these extra information sources can be either OOD or ID.

Moreover, Internet search engines are common approaches for humans to obtain external knowledge Komeili et al. (2021). More research opportunities can be explored to build Internet-augmented OOD detectors that can utilize rich and updated knowledge yielded by search engines to enhance the OOD detection performance.

OOD Detection and Lifelong Learning All previous approaches focus on detecting OOD inputs so that we can safely ignore them. However, OOD inputs usually represent new tasks that the current system does not support. Systems deployed in an ever-evolving environment are usually expected to continuously learn from these OOD inputs (without a full re-training) rather than ignoring them Liu & Mazumder (2021). However, humans exhibit outstanding abilities in learning new tasks from OOD inputs. We believe OOD detectors are essential components in a lifelong learning system, and it is helpful to combine OOD detection with a downstream lifelong learning process to build stronger systems. Specifically, a possible scenario is to present a subset of detected OOD samples to human annotators and apply a lifelong learning algorithm to absorb these annotations without re-training the original model. Further works can be carried out to integrate these processes to pursue more human-like AI systems (Kim et al., 2022; He & Zhu, 2022).

Theoretical Analysis of OOD Detection Despite impressive empirical results that OOD studies have achieved, theoretical investigation of OOD detection is far behind the empirical success (Morteza & Li, 2022; Fang et al., 2022). We hope more attention can be paid to theoretical analysis for OOD detection and provide insights to guide the development of better algorithms and applications.

7 Conclusion

In this survey, we provide a comprehensive review of OOD detection methods in NLP. We formalize the OOD detection tasks and identify the major challenges of OOD detection in NLP. A taxonomy of existing OOD detection methods is also provided. We hope this survey helps researchers locate their target problems and find the most suitable datasets, metrics, and baselines. Moreover, we also provide some promising directions that can inspire future research and exploration.

Limitations

There are several limitations of this work. First, this survey mainly focuses on OOD detection approaches for NLP domains. Despite the restrictive scope, our work well complements the existing survey on OOD detection in CV tasks, and hence will benefit a well-targeted research community in NLP. Second, some

OOD detection methods mentioned in this paper are not extended in this survey due to space limitations. We include details that are necessary to outline the development of OOD detection methods so that readers can get a comprehensive overview of this field. Our survey provides an elaborate starting point for readers who want to dive deep into OOD detection for NLP. Moreover, The term "OOD detection" has various alias, such as "Anomaly Detection", "Outlier Detection", "One-class Classification", "Novelty Detection", and "Open Set Recognition". These notations represent similar tasks with subtle differences in detailed experiment settings. We do not extensively discuss these differences due to space limitations. Readers can refer to other papers for more detailed discussions (Yang et al., 2021). Finally, we do not present any new empirical results. It would be helpful to perform comparative experiments over different OOD detection methods (Yang et al., 2022b). We leave this to future work.

Ethics Statement

This work does not present any direct ethical issues. In this survey, we provide a comprehensive review of OOD detection methods in NLP, and we believe this study leads to intellectual merits that benefit from a reliable application of NLU models.

References

Shikha Agrawal and Jitendra Agrawal. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713, 2015.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pp. 1454–1471. PMLR, 2023.

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1893–1902, 2015.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Christophe Bertero, Matthieu Roy, Carla Sauvanaud, and Gilles Trédan. Experience report: Log mining using natural language processing and application to anomaly detection. In *2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 351–360. IEEE, 2017.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.

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. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Terrance E Boult, Steve Cruz, Akshay Raj Dhamija, Manuel Gunther, James Henrydoss, and Walter J Scheirer. Learning and the unknown: Surveying steps toward open world recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9801–9807, 2019.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Derek Chen and Zhou Yu. Gold: Improving out-of-scope detection in dialogues using data augmentation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 429–442, 2021.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1657–1668, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

DongHyun Choi, Myeong Cheol Shin, EungGyun Kim, and Dong Ryeol Shin. Outflip: Generating examples for unknown intent detection with natural language attack. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 504–512, 2021.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023a.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring large language models for multimodal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*, 2023b.

Soham Dan and Dan Roth. On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2096–2101, 2021.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022b.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

Geli Fei and Bing Liu. 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*, pp. 506–514, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Varun Gangal, Abhinav Arora, Arash Einolghozati, and Sonal Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7764–7771, 2020.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

Jiangpeng He and Fengqing Zhu. Out-of-distribution detection in unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3850–3855, 2022.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2744–2751, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Gyuhak Kim, Sepideh Esmaeilpour, Changnan Xiao, and Bing Liu. Continual learning based on ood detection and task masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3856–3866, 2022.

Joo-Kyung Kim and Young-Bum Kim. Joint learning of domain classification and out-of-domain detection with dynamic class weighting for satisficing false acceptance rates. *arXiv preprint arXiv:1807.00072*, 2018.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in-and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1326–1340, 2020.

Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Hao Lang, Yinhe Zheng, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. Estimating soft labels for out-of-domain intent detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 261–276, 2022.

Hao Lang, Yinhe Zheng, Binyuan Hui, Fei Huang, and Yongbin Li. Out-of-domain intent detection considering multi-turn dialogue contexts. *arXiv preprint arXiv:2305.03237*, 2023.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. 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)*, pp. 1311–1316, 2019.

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

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018b.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pp. 587–592, 2003.

Xiaoya Li, Jiwei Li, Xiaofei Sun, Chun Fan, Tianwei Zhang, Fei Wu, Yuxian Meng, and Jun Zhang. kfolden: k-fold ensemble for out-of-distribution detection-fold ensemble for out-of-distribution detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3115, 2021.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Ting-En Lin and Hua Xu. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5491–5496, 2019.

Bing Liu and Sahisnu Mazumder. Lifelong and continual learning dialogue systems: learning during conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15058–15063, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

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

Denis Lukovnikov, Sina Daubener, and Asja Fischer. Detecting compositionally out-of-distribution examples in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 591–598, 2021.

Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

Petr Marek, Vishal Ishwar Naik, Anuj Goyal, and Vincent Auvray. Oodgan: Generative adversarial network for out-of-domain data generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 238–245, 2021.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=jlAjNL8z5cs`.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *arXiv preprint arXiv:2211.13445*, 2022.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Johannes EM Mosig, Shikib Mehri, and Thomas Kober. Star: A schema-guided dialog dataset for transfer learning. *arXiv preprint arXiv:2010.11853*, 2020.

Armineh Nourbakhsh and Grace Bang. A framework for anomaly detection using language modeling, and its applications to finance. *arXiv preprint arXiv:1908.09156*, 2019.

Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. Energy-based unknown intent detection with data manipulation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2852–2861, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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)*, pp. 2463–2473, 2019.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13675–13682, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6838–6855, 2020.

Mrinal Rawat, Ramya Hebbalaguppe, and Lovekesh Vig. Pnpood: Out-of-distribution detection for text classification via plug andplay data augmentation. *arXiv preprint arXiv:2111.00506*, 2021.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*, 2022.

Andras Rozsa, Manuel Günther, and Terrance E Boult. Adversarial robustness: Softmax versus openmax. *arXiv preprint arXiv:1708.01697*, 2017.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In Anoop Sarkar and Michael Strube (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5004. URL `https://aclanthology.org/N19-5004`.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32, 2017.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 714–718, 2018.

Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3795–3805, 2019.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. Enhancing the generalization for intent classification and out-of-domain detection in slu. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2443–2453, 2021.

Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2911–2916, 2017.

Lei Shu, Yassine Benajiba, Saab Mansour, and Yi Zhang. Odist: Open world classification via distributionally shifted instances. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3751–3756, 2021.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*, 2022.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. Out-of-domain detection for low-resource text classification tasks. 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)*, pp. 3566–3572, 2019.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1995–2002, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.158. URL `https://aclanthology.org/2022.findings-acl.158`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 550–564, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

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

Yanan Wu, Keqing He, Yuanmeng Yan, QiXiang Gao, Zhiyuan Zeng, Fujia Zheng, Lulu Zhao, Huixing Jiang, Wei Wu, and Weiran Xu. Revisit overconfidence for OOD detection: Reassigned contrastive learning with adaptive class-dependent threshold. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4165–4179, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.307. URL `https://aclanthology.org/2022.naacl-main.307`.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, 2021.

Hu Xu, Bing Liu, Lei Shu, and P Yu. Open-world learning and application to product classification. In *The World Wide Web Conference*, pp. 3413–3419, 2019.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 62–69, 2015.

Keyang Xu, Tongzheng Ren, Shikun Zhang, Yihao Feng, and Caiming Xiong. Unsupervised out-of-domain detection via pre-trained transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1052–1061, 2021.

Rakesh Bahadur Yadav, P Santosh Kumar, and Sunita Vikrant Dhavale. A survey on log anomaly detection using deep learning. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1215–1220. IEEE, 2020.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS Lam. Unknown intent detection using gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1050–1060, 2020.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022a.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022b.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. 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)*, pp. 870–878, 2021a.

Zhiyuan Zeng, Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. Adversarial generative distance-based classifier for robust out-of-domain detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7658–7662. IEEE, 2021b.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3521–3532, 2021.

Bangzuo Zhang and Wanli Zuo. Learning from positive and unlabeled examples: A survey. In *2008 International Symposiums on Information Processing*, pp. 650–654. IEEE, 2008.

Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. In *AAAI*, pp. 14374–14382, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209, 2020.

Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2021a.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1100–1111, 2021b.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. KNN-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5129–5141, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.352. URL `https://aclanthology.org/2022.acl-long.352`.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

A Surveying Process

In this appendix, we provide more details of how we select papers for our survey. Specifically, the selected paper follows at least one criterion listed below:

1. Peer-reviewed papers published in Top-tier NLP venues, such as ACL, EMNLP, NAACL, AAAI, and IJCAI.
2. Peer-reviewed papers that have a significant impact on the OOD detection area. These papers are not necessarily limited to NLP tasks.
3. Papers that are highly cited in the OOD detection area.
4. Most recently published papers that make a non-trivial contribution to OOD detection, such as methods, datasets, metrics, and theoretical analysis.
5. Papers that initiate each research direction in the OOD detection area.

B More details of Datasets

Table 1 provides more detailed information of various common datasets for OOD detection, regarding the total number of classes, the data size of ID and OOD samples, and selected papers using these datasets.

C More details of Metrics

Table 2 provides more detailed information of various metrics for OOD detection, regarding whether to consider ID performance, frequency of use, and applications.

Dataset	Classes	#ID	#OOD	Papers that use this dataset (Selected)
CLINC150 (Larson et al., 2019)	150	22,500	1,200	(Zhang et al., 2021; Zhan et al., 2021; Lang et al., 2022; Zhou et al., 2022)
Banking (Casanueva et al., 2020)	77	13,083	0	(Zhang et al., 2021; Zhan et al., 2021; Lang et al., 2022; Zhou et al., 2022)
StackOverflow (Xu et al., 2015)	22	20,000	0	(Zhang et al., 2021; Zhan et al., 2021; Lang et al., 2022; Zhou et al., 2022)
STAR (Mosig et al., 2020)	150	27,510	1,594	(Chen & Yu, 2021; Lang et al., 2023)
ROSTD (Gangal et al., 2020)	12	43,323	4,590	(Chen & Yu, 2021; Podolskiy et al., 2021)

Table 1: More detailed information of various common datasets for OOD detection. # indicates the total number of samples.

Metric	Definition	Whether to consider ID performance	Frequency of use	Applications	Papers that use this metric (Selected)
AUROC	Area under the Receiver Operating Characteristic curve	No	Very Frequent	NLP, CV, ML	(Hendrycks & Gimpel, 2016; Hendrycks et al., 2018; 2019; Lee et al., 2018a)
AUPR-IN	Area under the Precision-Recall curve (ID samples as positive)	No	Frequent	NLP, CV, ML	(Lee et al., 2018a; Zheng et al., 2020; Shen et al., 2021)
AUPR-OUT	Area under the Precision-Recall curve (OOD samples as positive)	No	Frequent	NLP, CV, ML	(Lee et al., 2018a; Zheng et al., 2020; Shen et al., 2021)
FPR@N	Value of FPR when TPR is N%	No	Not Frequent	NLP, CV, ML	(Lee et al., 2018a; Zheng et al., 2020; Shen et al., 2021)
F1	Macro F1 score over all testing samples (ID+OOD)	Yes	Very Frequent	NLP	(Xu et al., 2019; Zhan et al., 2021; Shu et al., 2021; Zhou et al., 2022)
Acc	Accuracy score over all testing samples (ID+OOD)	Yes	Very Frequent	NLP	(Zhan et al., 2021; Shu et al., 2017; 2021; Zhou et al., 2022)

Table 2: More detailed information of various metrics for OOD detection.