

A Review of Out-of-Distribution Generalization in Text Classification Models

Ximing Yao*

School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

**Corresponding author: Ximing Yao.*

Abstract

A basic premise of standard machine learning is that the training set and the test set come from the same distribution, but this premise is often broken in practice. Text classification is one of the most fundamental tasks in natural language processing (NLP). The poor robustness of its models to out-of-distribution (OOD) samples has become an important research topic. This study reviews the latest research on OOD generalization in text classification. First, it defines OOD generalization, which can be viewed from the data perspective as systematic data variance and from the model perspective as distorted features. Second, it introduces domain generalization methods, anti-shortcut feature learning, solutions for class distribution shift, and the main methods for OOD generalization of large language models. Then, it presents evaluation benchmarks and datasets used for three types of evaluation: distribution shift, compositional generalization, and class distribution shift. Finally, it identifies the limitations of existing studies and proposes future research directions.

Keywords

out-of-distribution generalization, text classification, domain generalization, robustness

1. Introduction

Natural language processing (NLP) systems face a basic problem in practical applications: a model performs well on the training set but cannot generalize to unseen data, that is, the data in the test set differ from those in the training set. Pretrained language models have reached or even exceeded human performance on common evaluation benchmarks such as sentiment analysis and text classification, but this high accuracy may be overestimated because the models rely excessively on certain systematic biases. Text classification is a fundamental NLP task and is widely used in spam filtering, email classification, sentiment tendency judgment, and other applications. However, real-world text data are highly diverse and full of uncertain factors, and new hot topics constantly emerge. If a classifier is trained under a closed-world setting, new classes that do not appear in the training set may appear during practical application. The classifier will classify these new classes as old classes, thereby reducing the accuracy of the whole system.

In recent years, out-of-distribution (OOD) generalization has become a hot research topic in text classification. In terms of methods, Ganin et al. proposed domain-adversarial training, which uses a gradient reversal layer to learn domain-invariant representations [1]. Li et al. proposed the meta-learning method MLDG to improve generalization ability during cross-domain training and testing [2]. Xie et al. proposed unsupervised data augmentation, which uses back-translation to generate diverse training samples [3]. In terms

of evaluation, Yang et al. created the GLUE-X benchmark, which includes 13 datasets, 8 types of tasks, and 21 pretrained models, showing that the average accuracy of cross-domain evaluation is far below human performance [4]. However, current research in this direction mainly focuses on robustness to adversarial attacks and causal explanation. Sebastiani summarized machine learning methods in text classification at an early stage [5]. Wang et al. analyzed model generalization from the perspective of domain generalization (DG) in their survey, but did not distinguish data-level variance from distorted features [6]. Yang et al. argued that previous studies lacked a summary of text classification from the OOD perspective and first proposed a classification framework based on “systematic data variance” and “distorted features” [7]. Costache et al. divided existing methods for the special case of class distribution shift into three types according to their settings: Universum learning, zero-shot learning, and open-set learning [8].

Current large-scale OOD generalization studies mainly summarize the relevant literature by “method classification”, but this approach has two limitations. First, it cannot distinguish whether generalization failure is caused by systematic bias in the data or by bias inherent in the model itself, so the understanding of the causes of generalization failure remains superficial. Second, it lacks a systematic review of the special OOD case of class distribution change, which is one of the most common distribution changes in practical text classification tasks. To address the above problems, this study summarizes the relevant literature from 2018 to 2026 and, on the basis of the “systematic data variance” and “distorted features” proposed by Yang et al., divides current work into two aspects: (1) handling systematic data variance at the data level, including DG methods and specific solutions for class distribution shift; and (2) handling distorted features at the model level, including anti-shortcut feature learning and OOD generalization ability of large language models. In addition, this study introduces the development of evaluation standards such as GLUE-X and BOLT. Compared with existing survey articles, this study not only provides a relatively comprehensive summary of OOD generalization methods and evaluation standards for text classification, but also focuses on targeted measures adopted for class distribution shift, the most common scenario in practical applications, as well as new challenges and opportunities brought to OOD generalization by the rise of large language models. Research in this area helps to understand from multiple perspectives why OOD generalization problems occur in text classification systems and provides useful reference for improving the robustness of text classification systems in production environments.

2. Concepts and Classification Framework of Out-of-Distribution Generalization

2.1 Problem Formalization

Assume that training samples are drawn from one distribution, while test samples come from another different distribution. In this case, an out-of-distribution problem is formed.

From a probabilistic perspective, the Bayesian method can be used to obtain the OOD detection function [7]. Let the model parameters be given, and let the model’s conditional OOD probability for a given input be defined. Then the probability that a sample is OOD can be obtained through model averaging:

$$P(\text{OOD} | x) = \sum_{\theta} P(\text{OOD} | \theta, x) P(\theta | x)$$

The OOD detection function can be defined as a threshold judgment on prediction confidence:

$$g(x) = \left[\max_y P(y | x) < \epsilon \right]$$

Where the confidence threshold parameter represents the prediction confidence level. When the model’s prediction confidence for an OOD sample is lower than this threshold, the sample is regarded as OOD [7].

It should be noted that this formalization is a necessary conceptual tool for clarifying the scope of the OOD detection problem and for classifying subsequent methods. In Section 3, DG does not require this detector; instead, it learns domain invariance during training to improve generalization performance. Open-set learning is based on the idea of threshold judgment and excludes unknown-class samples while classifying known classes [8].

2.2 Classification Framework

OOD generalization mainly consists of two parts: systematic data variance and distorted features [7]. Systematic data variance refers to changes viewed from the perspective of the data itself, such as language characteristics in different domains or different subgroups. DG addresses this problem by making predictions in one or more unseen new domains, which do not appear during training. These changes are generated by the data itself rather than by model learning [6]. Distorted features are misleading associations learned by a model. During training, the model uses shortcut features in the training set to make predictions instead of learning semantic features with real causal relationships. For example, in a natural language inference task, if longer sentences correspond to larger labels, the model will make judgments according to this pattern rather than perform effective reasoning [7].

According to the specific form of shift, OOD scenarios can be further divided into the following types, as shown in Table 1:

Table 1: Types of Out-of-Distribution Scenarios

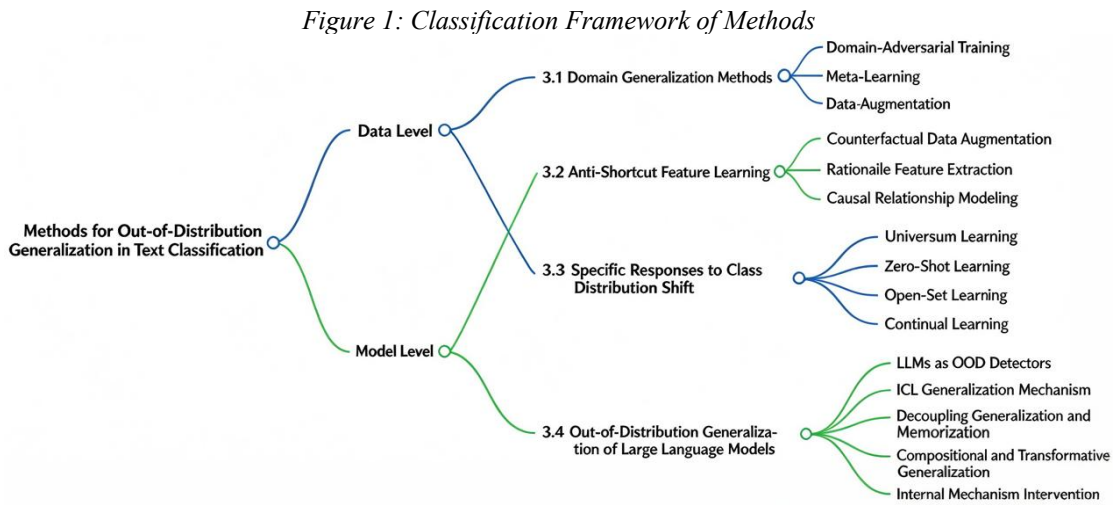
Shift Type	Description	Typical Example
Domain Shift	The data source domain changes.	Training on news data and testing on social media text.
Class Shift	New classes appear during testing.	New topic classes appear after training.
Temporal Shift	Data evolve over time.	Changes in news topics before and after COVID-19.
Language Shift	Cross-lingual scenarios.	Training on English data and testing on Chinese data.

Among them, class distribution shift is a situation that requires attention. In an open environment, there are “unknown unknowns”, that is, new classes that the model has never seen. According to different levels of class knowledge, it can be further divided into three types.

- Universum learning: There are unlabeled background-class samples, and samples that do not belong to any known class are assigned to the background class.
- Zero-shot learning: The existence of known classes is known, but there are no training samples. On this basis, new classes are inferred through class semantics.
- Open-set learning: The existence of unknown classes is completely unknown during training, so the model needs to detect and discover the emergence of new classes [8].

3. Main Methods and Technical Routes

OOD generalization methods for text classification can be summarized from the data perspective and the model perspective, as shown in Figure 1:



3.1 Domain Generalization Methods

The goal of DG is to obtain a model that can perform well in newly emerging target domains. Wang et al. [6] summarized existing DG methods into three major categories: data manipulation, representation learning, and learning strategy. Yang et al. [7] further reviewed the development of OOD generalization from traditional machine learning to deep learning and then to large language models (LLMs). Methodologically, DG can be seen as developing from “explicit alignment” to “implicit learning”.

Domain-adversarial training is an explicit alignment method. The domain-adversarial neural network (DANN) proposed by Ganin et al. [1] uses a gradient reversal layer to enable the feature extractor to learn domain-invariant representations. Its basic assumption is that “domain information can be removed from features”. However, this assumption is not always feasible in text classification. Content differences caused by different language styles are difficult to separate, and excessive alignment can affect class discrimination ability. Muandet et al. [9] first proposed the idea of obtaining domain-invariant representations by minimizing representation differences among source domains. Recent studies have combined adversarial training with active learning. The DAAL algorithm improves performance by selecting domain-adversarial samples and achieves an improvement of more than 10% using only 4% of the data [10].

Different from methods that remove domain information through adversarial training, invariant learning aims to find an effective prediction method that can be applied in different environments. Arjovsky et al. [11] proposed invariant risk minimization (IRM), which transforms the generalization problem into finding a data representation under which there exists a classifier whose loss is minimized in every training environment. Krueger et al. [12] proposed V-REx, which adds a risk variance penalty term, requiring the risks of the model in different environments not only to be small but also to be as close as possible. Sagawa et al. [13] proposed GroupDRO from the perspective of distributional robustness, attempting to minimize group risk in the worst case. Sun and Saenko [14] proposed CORAL, which achieves distribution matching by matching the second-order moments of features across domains. Their common point is that they all assume a causal invariance that runs through all environments, while shortcut features do not work across different environments.

The meta-learning route provides a new perspective. The MLDG proposed by Li et al. [2] allows the model to learn “how to quickly adapt to a new domain” by imitating cross-domain training and testing, instead of trying to perform best in one domain. The implication behind this is that generalization does not mean “removing differences between domains”, but rather “obtaining the ability to adapt”. Balaji et al. [15] further applied this idea to adaptive regularization in feature space through MASF.

Data augmentation methods mitigate this variance problem from the input side. The unsupervised data augmentation (UDA) method proposed by Xie et al. [3] uses back-translation to generate samples with the same semantics but different appearances. This is based on the premise of “semantic invariance”, that is, class labels do not change because the appearance changes. Easy data augmentation (EDA) proposed by Wei and Zou [16] includes four operations: synonym replacement, random insertion, random swap, and random deletion. In essence, it perturbs data at the word level. However, it should be pointed out that these methods perform differently across tasks. In text classification, they have significant effects on topic classification, but sentiment classification is very sensitive to synonym replacement.

3.2 Anti-Shortcut Feature Learning

Distorted features refer to erroneous associations learned by a model: the model uses shortcut features in training data, such as certain words or sentence length, rather than truly meaningful semantic features to make predictions. Methodologically, anti-shortcut feature learning develops from “data intervention” to “structural modeling”.

Counterfactual data augmentation is a data intervention method. Kaushik et al. [17] proposed a method of counterfactually augmented data at ICLR, asking human annotators to modify documents so that they fit counterfactual target labels while maintaining internal consistency. An important finding of this method is that simply adding more data cannot solve the problem of spurious correlations. It is also necessary to use “counterfactual editing” to break the shortcut between features and labels. Wang and Culotta [18] proposed a robust text classification method based on automatically generated counterfactuals at AAAI, enabling the classifier to obtain more reliable predictions by focusing on causal features and ignoring non-causal features.

Wu et al. [19] proposed a counterfactual data augmentation method for aspect-based sentiment analysis at ACML. Zhang et al. [20] proposed a counterfactual data augmentation method based on dual self-improvement at ACL, using the attention of the task model to find causal entities related to the task. A common point of these studies is the transition from using manually annotated counterfactual samples to using signals from the model itself for counterfactual sample generation.

Rationale extraction is a form of structural modeling. Lei et al. [21] proposed Rationalizing Neural Predictions at EMNLP, using a “generator-encoder” method to obtain concise and fluent rationale fragments, which performed better than attention-based methods on multiple sentiment tasks. Chang et al. [22] proposed invariant rationalization at ICML, aiming to identify rationale fragments that are consistent across domains and thereby improve the robustness of the model to unknown data. This study raised rationalization from an “interpretation method” to a “generalization technique”. Yu et al. [23] proposed the A2R framework at NeurIPS, using an additional predictor based on a soft attention mechanism during rationalization to solve the “interlocking dynamics” between the generator and the encoder in the rationalization process.

Causal relationship modeling advances intervention to the structural level, that is, from the feature level to the structural level. Zhou and Zhu [24] proposed a causally calibrated robust classifier and reported at NAACL that counterfactual reasoning can be used to obtain the importance and sufficiency probability of each feature, which are then combined with an unbiased inverse-propensity-weighted loss to suppress spurious correlations without group labels. Veitch et al. [25] proposed counterfactual invariance learning, which discovers causal features through counterfactual invariance constraints. The development from counterfactual augmentation to causal modeling shows that the greatest difficulty in anti-shortcut learning is not “finding which features are shortcuts”, but “finding a method that can automatically separate causal features without group labels”.

3.3 Specific Responses to Class Distribution Shift

Class distribution shift is the most common OOD situation in text classification. New topics constantly emerge, and in production there are often classes that have not been seen before. Different from domain shift, which refers to changes in data sources, class shift means that the classes themselves change. Therefore, traditional domain adaptation methods are not necessarily effective. According to different degrees of class knowledge, it can be divided into three situations.

Universum assumes that there are many unlabeled background-class samples and uses these samples to learn a background-class classifier, thereby absorbing samples that “do not belong to any known class”. Universum learning was first proposed by Weston et al. [26], who added unlabeled background samples into support vector machines (SVMs) as a regularization term. However, one problem with Universum learning is that it assumes background-class samples can be obtained, while in reality new-class samples generally appear only after the model is deployed and cannot be obtained in advance.

Zero-shot learning (ZSL) uses class semantic descriptions, such as textual definitions or attribute vectors, as a bridge connecting known classes to unknown classes. Palatucci et al. [27] first applied ZSL to neuroimaging data. In NLP, Pushp and Srivastava [28] used the semantic embedding of each class name for zero-shot text classification. The performance of ZSL largely depends on the quality of class semantic descriptions. If new classes and known classes differ greatly in semantics, it is difficult to build this bridge.

Open-set learning requires the model to correctly identify known classes while excluding samples of unknown classes. Scheirer et al. [29] proposed a method based on compact abating probability, namely W-SVM, to reduce open-set risk to a finite level. Bendale and Boulton [30] proposed the nearest non-outlier method, which uses distance to detect newly added classes one by one. For text classification, the BOLT benchmark [31] includes two tasks, open-set text classification and generalized category discovery. After comparing 22 methods, it found that most methods tend to overfit the training set and cannot generalize well to new classes.

Continual learning is also a way to solve class distribution shift. Costache et al. [8] concluded that continual learning can effectively handle class distribution shift. Continual learning adopts a “stability-plasticity” strategy so that the model does not forget existing knowledge while learning new classes. For example, iCaRL proposed by Rebuffi et al. [32] and LwF proposed by Li and Hoi [33] are such methods. In addition, elastic weight consolidation proposed by Kirkpatrick et al. [34] uses a small regularization term to restrict changes to important parameters and has achieved good performance on multiple tasks and incremental text classification.

From Universum to zero-shot learning and then to open-set learning, the development of technical routes reflects a contradictory issue: how information from known classes can help unknown classes. Zero-shot learning uses semantic priors, open-set learning uses distribution boundaries, and continual learning seeks a compromise between the two.

3.4 Out-of-Distribution Generalization of Large Language Models

The emergence of large language models has not solved the OOD problem, and has even made the problem appear in a new form. Different from the traditional sense of “inconsistency between the training set and the test set”, the OOD problem of LLMs is caused by limitations in their in-context learning ability during modeling. Current research mainly has two directions: one is to use LLMs as tools for OOD detection, and the other is to study the OOD ability of LLMs themselves.

From the perspective of detection tools, Xu and Ding [35] first published a systematic survey in Findings of NAACL. According to the different roles played by LLMs, they divided existing work into two types: one type uses LLMs as zero-shot OOD detectors, which judge whether an example comes from outside the training set only through their own understanding of language without any additional information; the other type uses LLMs as auxiliary modules, such as assisting in generating pseudo-labels or providing semantic embeddings, to improve the effect of the original detector.

From the perspective of generalization, Kwon et al. [36] pointed out that in-context learning (ICL) generalization ability is determined by the subspace of pretraining data. When the task distribution can be represented as the sum of several connected subspaces, the model can generalize to any subspace in the space spanned by the training subspaces, even if the probability of this subspace in the training distribution is 0. When the training data lie only in one subspace, inference can only be carried out in this subspace and cannot generalize outside it. For text classification, this indicates a problem: the “zero-shot generalization” shown by LLMs in cross-domain classification may actually be an illusion, because the pretraining data may already contain the potential structure of the target domain, that is, the test task vector is located in the space spanned by the training task vectors. If the target domain, such as low-resource languages or specialized legal and regulatory documents, is not in the space formed by the pretraining data, the generalization ability of LLMs will decline substantially.

The Scylla evaluation method proposed by Qi et al. [37] also points out the relationship between generalization ability and memorization. Through separation measures across 20 tasks and 5 complexity levels, it finds that the more difficult the task is, the larger the in-distribution (ID)-OOD performance gap becomes. There is a critical point before which the model’s reliance on non-generalizing behavior gradually increases. An implication for text classification is that when measuring the classification performance of LLMs, it is necessary to distinguish whether “the task itself is difficult” or “the model is memorizing”. For tasks close to this critical point, good LLM performance may be an overestimation of its actual generalization ability.

The OMEGA benchmark proposed by Sun et al. [38] divides OOD generalization into exploratory generalization, compositional generalization, and transformative generalization. Its results show that LLMs still need improvement in compositional generalization. This also indicates a problem for text classification: most current evaluations focus only on “cross-domain” settings and ignore a possible type of compositional generalization in classification tasks, for example, a new class composed of several known classes. Suresh et al. [39] used sparse autoencoders to track the internal activation of Transformers and found that OOD samples cause the model to produce more spurious concepts. They accordingly proposed a simple method for fine-tuning. This shows that in the LLM era, research on OOD generalization is shifting from “detection and classification” to “diagnosis and intervention of internal mechanisms”, and text classification tasks also need to find their own position in this process.

4. Evaluation Benchmarks and Datasets

Evaluation benchmarks are basic tools for measuring the development of OOD generalization research. In text classification, three types of evaluation benchmarks have emerged: integrated evaluation for distribution shift, specialized benchmarks focused on compositional generalization, and open-world benchmarks for class distribution shift. The following summarizes these three types of benchmarks.

4.1 Distribution Shift Benchmarks

GLUE-X is currently the most influential OOD generalization evaluation standard in text classification. Yang et al. [4] proposed the first unified evaluation standard for OOD robustness, including 13 public datasets, 8 classic NLP tasks, and 21 mainstream pretrained language models. The most important finding of this standard is that even the best model has an average accuracy in cross-domain testing far lower than the human level (74.6% vs. 80.1%), which stands in sharp contrast to the fact that more than 20 single-model results on the GLUE leaderboard are better than human performance. This also shows that the prosperity under standard GLUE evaluation is only superficial, and high scores within distribution do not mean that a model has good generalization ability.

For document image classification, ZOD is the first dataset supporting zero-shot learning and OOD detection. Sheikh et al. [40] created a dataset at the ICCV Workshop that contains approximately 38,000 images and 10 classes, of which 5 classes are the same as those in the traditional RVL-CDIP and the other 5 classes are new. This fills the gap of lacking a standard benchmark in document images that can simultaneously support ZSL and OOD detection, and provides a unified standard for testing the robustness of document classifiers.

In addition, for OOD detection in text classification, some scholars have proposed benchmark datasets including non-semantic shift, such as Yahoo-AGNews-five, and semantic shift, such as AGNews-FL and AGNews-FM [41]. Yahoo-AGNews-five is constructed by selecting part of Yahoo! Answers and the AG Corpus to test the recognition effect of the model when label semantics are consistent but sources differ. AGNews-FL and AGNews-FM use OOD classes with larger semantic differences to create scenarios with different degrees of semantic shift. Systematic benchmark testing of classic OOD detection methods also uses three sentiment classification datasets, IMDB, SST-2, and Yelp, and on this basis forms different ID/OOD pairings, including Near-OOD, Far-OOD, and Distinct-OOD [42]. A comparison of distribution shift benchmarks is shown in Table 2:

Table 2: Comparison of Distribution Shift Benchmarks

Benchmark	Domain	Scale	Task	Key Finding
GLUE-X	General NLP	13 datasets	OOD generalization evaluation	Cross-domain accuracy: 74.6% vs. 80.1% human.
ZOD	Document images	38,000 images	ZSL; OOD detection	Supports ZSL and OOD detection.
Yahoo-AGNews-five	News; Q&A	675K train	Nonsemantic shift	OOD detection benchmark.
AGNews-FL / AGNews-FM	News text	116K train	Semantic shift	Evaluates semantic shift.
IMDB / SST-2 / Yelp	Sentiment analysis	25K / 67K / 560K	Classic OOD detection	Near-, Far-, and Distinct-OOD.

4.2 Compositional Generalization Benchmarks

Compositional generalization benchmarks are important standards for measuring a model’s ability to generalize to “new combinations of known components”. The SCAN dataset proposed by Lake and Baroni [43] first comprehensively showed the shortcomings of sequence-to-sequence models in recombining known primitives to generate new commands, and it is the first benchmark for compositional generalization. In addition, COGS proposed by Kim and Linzen [44] uses a semantic parsing task in which a model transforms a sentence into the corresponding logical form to examine lexical generalization and structural generalization. CFQ proposed by Keyzers et al. [45] uses maximum compound divergence (MCD) to make the combinations in the training set and the test set very different, and it is currently one of the most difficult compositional generalization benchmarks.

Compositional generalization failure is systematic. The problem first found by Lake and Baroni [43] on SCAN was repeatedly reproduced later on COGS [44], CFQ [45], and other benchmarks. Some review articles mention that this gap consistently exists across different model families. Although larger scale can alleviate part of the gap, it cannot eliminate it, while structured decoding is most effective in reducing the gap [46]. This indicates that compositional generalization ability cannot be naturally obtained simply by increasing model parameters.

4.3 Class Distribution Shift Benchmarks

Class distribution shift benchmarks focus on situations in which the model encounters new classes during testing that have never been seen during training. MOSLD-Bench is the first multilingual open-set learning and discovery benchmark for text classification. It includes 12 different languages, 960K data points, and all data come from the news domain. This benchmark is more demanding because the new classes are completely unknown during training, which requires the model to discover and learn these new classes by itself. Costache et al. [47] proposed the OSLD method, which uses clustering and contrastive learning to discover classes and uses pretrained language models to compare semantic similarity for label alignment, achieving 84.5% accuracy on Arabic T1.

BOLT is a unified benchmark and evaluation toolkit for open-world text classification, including two main tasks. Open-set text classification (OSTC) requires the model to classify examples belonging to training-set classes and reject OOD inputs. Generalized category discovery (GCD) requires the model to identify both known and new classes when only part of the data is labeled. Qin et al. [31] conducted a systematic comparative study at ACL 2026 on 12 public datasets and 22 methods, including 15 OSTC methods and 7 GCD methods, showing that most current methods overfit the training set and cannot generalize well to unseen classes. By comparing lightweight LLM variants with existing open-set baselines, this work demonstrates the possibility of using LLMs for open-world text classification.

For document images, although RVL-CDIP is the “de facto benchmark” for document classification, the study by Larson et al. [48] pointed out that its annotation errors account for about 12%, and about 35% of the data appear in both the test set and the training set. After training based on cleaned data, the OOD generalization accuracy of supervised models increased by an average of 8.1 percentage points, with the highest increase reaching 14 percentage points, indicating that data quality is also one of the factors affecting OOD generalization.

5. Challenges and Prospects

5.1 Current Challenges

Inconsistent evaluation standards: Because there is no unified understanding of “what out-of-distribution data are”, the academic community differs in its understanding of “what degree of deviation counts as out-of-distribution”. This leads to confusion among concepts such as OOD detection, OOD generalization, and out-of-distribution generalization. Although these terms differ in name, they are actually caused by different understandings of problem definition, measurement standards, and solutions. OOD detection is aimed at “detecting out-of-distribution samples”, while OOD generalization means “making good predictions on out-of-distribution samples”. The two differ greatly in problem setting, evaluation standards, and solution methods, but many papers treat them as equivalent, making it difficult for readers to distinguish their real scope of application. In addition, there is still no clear standard for “what degree of distribution shift counts as OOD”. In images, researchers can measure the degree of deviation by the magnitude of pixel-level perturbation. In text, however, there is no unified standard for measuring semantic shift itself. If one document changes its content while retaining sentence structure, and another document retains content but changes style, which one has a larger shift? There is no unified understanding of such problems, resulting in the inability to make effective comparisons among different studies.

Insufficient evaluation: Existing studies usually examine only one aspect of distribution shift and ignore model performance on different types of OOD. Yang et al. pointed out that although current work has achieved certain results, there is still much room for improvement in generalization ability. Specifically, existing work has three kinds of separation. The first is separation in shift types: DG studies are generally evaluated only in cross-domain settings, compositional generalization studies are tested only on compositional generalization benchmarks, and studies of class shift mainly focus on open-set problems, with little overlap among the three aspects. The second is separation in evaluation metrics: different studies use different metrics; some report accuracy drop, some report worst-domain performance, and some report OOD detection AUROC, making cross-study comparison difficult. The third is separation in model scale: many OOD generalization studies still use small-scale models such as BERT-base, while large-scale language models have become common in practice, so whether these results can generalize remains unclear. These three separations make it impossible

to answer an important question: Will a technique that is effective for multiple types of tasks also be effective under class shift? Almost no related work has systematically explored this issue.

Disconnection between methods and evaluation: Different research directions such as DG, anti-interference learning, and compositional generalization are independent of each other and lack a general method for comparison. Therefore, it is impossible to comprehensively compare these different methods or learn lessons from them. The deeper reason is that these different research routes are based on inconsistent premises. DG assumes that multiple source domains are seen during training and that invariant parts are learned on this basis for generalization. Anti-shortcut learning assumes that the model can identify what is truly relevant, but needs to know which features are “irrelevant”. Compositional generalization assumes that the model will reorganize previously learned abilities to solve new problems. These assumptions are effective in their own experiments, but in practical applications they often exist simultaneously. OOD in the real world includes domain shift, shortcut features, and new combinations. At present, there is no method that can handle multiple shifts, resulting in poor “composability” among methods. In addition, different studies use different baselines, different dataset split methods, and even different hyperparameter search methods. Therefore, even studies based on the same dataset may have different results because of experimental settings rather than the effectiveness of the methods themselves.

5.2 Future Research Directions

From detection to adaptation: It is necessary not only to detect OOD samples, but also to enable models to learn new distributions. Continual learning has been shown to help alleviate class distribution shift, but more work is still needed in this area. At present, most OOD studies focus on “detection”, that is, judging whether a sample is out-of-distribution and, if it is, rejecting service or reducing service quality. However, a truly robust system should not only do this. It should also have the ability of “online adaptation”. When the model discovers that the distribution has changed, it can use a small amount of unlabeled or labeled data from the new distribution for self-adjustment rather than simply refusing service. Continual learning provides certain support for this direction, and its greatest problem is the “stability-plasticity dilemma”, that is, the model needs to learn new datasets but cannot forget previous knowledge. Combining OOD detection with continual learning to form a cycle of “detection-adaptation-consolidation” is a direction worth studying. In addition, how to judge adaptation effectiveness without any annotation information, namely the evaluation of unsupervised adaptation effectiveness, remains unsolved.

Formulation of unified evaluation standards: Establishing an integrated evaluation standard that includes different types of distribution shift can promote the development of OOD generalization. GLUE-X has been proposed, but it still needs further improvement. The progress of GLUE-X lies in its first attempt to standardize OOD evaluation, but it mainly focuses on English NLU tasks and lacks sufficient support for Chinese, multilingual, and generative tasks. Future unified benchmarks should be improved in the following aspects. First, the richness of shift types should be improved. The benchmark should include not only domain shift, class shift, temporal shift, and language shift, but also various combined shifts, such as simultaneous changes in domain and class. Second, evaluation standards should be consistent. On this basis, it is recommended to report the “mean”, “worst-case performance”, and “performance drop” at the same time to prevent misleading conclusions caused by a single metric. Third, models of different sizes should be studied on the same benchmark, ranging from small models such as TextCNN to LLMs of various scales, to reveal the real relationship between scale and generalization. Finally, dynamic changes should be included. A benchmark should not be a fixed test set, but should gradually increase in difficulty as model capability improves, so as to prevent leaderboard overfitting.

Research on the generalization of large models: It is necessary to analyze the relationship between scale effects and generalization, as well as the performance of large models under extreme distribution shift. “Emergent generalization of LLMs” needs further exploration. At present, there are two opposing views. One view holds that because LLMs have massive pretraining data and ultra-long context windows, they have better OOD generalization ability than traditional models and can even reach human level in some cases. The other view holds that even so, LLM generalization is still “scaled pattern matching” and remains very weak when true abstract thinking or causal reasoning is involved. Both views have some basis but lack sufficient experimental verification. Future research also needs to address several questions. First, whether LLM OOD generalization ability has an “emergence threshold”, that is, whether a sudden change in generalization ability

will occur above a certain number of parameters or amount of pretraining samples. Second, whether LLM in-context learning can completely replace traditional methods for domain adaptation. Third, when the distribution difference is very large, for example, from news articles to poetry or from modern Chinese to classical Chinese, where the generalization boundary of LLMs lies. These questions are not only technical choices, but also concern how understanding and memory are perceived.

6. Conclusion

This study provides an overview and review of work related to OOD generalization in text classification. OOD generalization is a bridge between theoretical research and application implementation. The main question considered in OOD generalization is whether a model can still achieve good classification performance when the test dataset distribution differs from the training set distribution. This problem is especially important in text classification because natural language is diverse, topics constantly change, and writing styles vary among individuals, making it difficult for the training distribution to match the actual use environment.

From the perspective of research progress, current work has made certain progress in some aspects. DG methods train models with domain-invariant features and outperform standard fine-tuning in cross-domain text classification tasks. Anti-shortcut learning methods reduce the model's dependence on surface patterns by eliminating unnecessary associations. Research on compositional generalization points out the insufficient robustness of neural networks to newly emerging combinations, thereby providing new ideas for judging whether a model truly understands. Class distribution shift is one of the OOD problems closest to practical application scenarios. At present, three relatively clear technical routes have been developed, including Universum learning, zero-shot learning, and open-set learning, and the idea of continual learning is also applicable to this field. In addition, the proposal of evaluation standards such as GLUE-X is a major progress in this field, marking a transition from "each study making its own claim" to "unified evaluation".

Overall, however, this field is still in its early stage and still faces some unresolved basic problems. First, concepts are unclear. Terms such as OOD detection, OOD generalization, and out-of-distribution generalization are frequently used in the literature, and the research scope is unclear, resulting in inefficient knowledge accumulation. Second, there are three separations in evaluation: there is a lack of cross-evaluation among different types of shifts, evaluation metrics lack consistency, and model sizes are not sufficiently comprehensive, making horizontal comparison among different studies difficult. Finally, there is a lack of unified theory in methods. DG, anti-shortcut learning, and compositional generalization are methods proposed based on different assumptions, and it is difficult to form an effective solution for multiple shifts. In addition, the closed loop from detection to adaptation has not yet been established. Most current work remains at the level of "detecting OOD samples" and lacks a method for real-time adaptation and continuous learning.

Theoretically, the two causes of OOD generalization, systematic data variance and distorted features, explain the reasons for this problem from the perspectives of data and model. Systematic data variance is generated because real-world data are inherently diverse, that is, "the world is more interesting than the training set". Distorted features arise because, during learning, the model learns content that it should not learn, that is, "the model learns what it should not learn". There is a relationship between the two: the larger the data variance is, the more easily the model learns distorted features; at the same time, the stronger the distorted features are, the more sensitive the model is to data variance. This coupling relationship shows that a solution from only one aspect cannot solve the problem fundamentally, and future research should seek breakthroughs in both data and models.

OOD generalization is not only a technical problem, but also a scientific problem concerning the "nature of generalization". Its solution requires coordinated and unified efforts in data, models, evaluation, and other aspects, and also requires drawing on new viewpoints and ideas from statistical learning theory, causal reasoning, and cognitive science. Only in this way can text classification systems move away from "achieving excellent results only on benchmark tests" and enter a stage in which they "can play a good role in practical applications".

References

- [1] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59), 1-35.
- [2] Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. (2018, April). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 32(1), 3490-3497.
- [3] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33, 6256-6268.
- [4] Yang, L., Zhang, S., Qin, L., Li, Y., Wang, Y., Liu, H., ... & Zhang, Y. (2023, July). Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the association for computational linguistics: ACL 2023*, 12731-12750.
- [5] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47.
- [6] Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... & Yu, P. S. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8), 8052-8072.
- [7] Yang, L., Song, Y., Ren, X., Lyu, C., Wang, Y., Liu, L., ... & Zhang, Y. (2023). Out-of-distribution generalization in text classification: Past, present, and future. *arXiv preprint arXiv:2305.14104*.
- [8] Costache, A. V., Gheorghe, S. F., Poesina, E., Irofti, P., & Ionescu, R. T. (2026, March). Text Classification Under Class Distribution Shift: A Survey. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4045-4060).
- [9] Muandet, K., Balduzzi, D., & Schölkopf, B. (2013, February). Domain generalization via invariant feature representation. In *International conference on machine learning* (pp. 10-18). PMLR.
- [10] Chen, J., Ding, L., Yang, Y., Di, Z., & Xiang, Y. (2024). Domain adversarial active learning for domain generalization classification. *IEEE Transactions on Knowledge and Data Engineering*, 37(1), 226-238.
- [11] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.
- [12] Krueger, D., Caballero, E., Jacobsen, J. H., Zhang, A., Binas, J., Zhang, D., ... & Courville, A. (2021, July). Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning* (pp. 5815-5826). PMLR.
- [13] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks. *International Conference on Learning Representations*.
- [14] Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. *European Conference on Computer Vision Workshops*, 443-450.
- [15] Balaji, Y., Sankaranarayanan, S., & Chellappa, R. (2018). MetaReg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31, 998-1008.
- [16] Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 6382-6388.
- [17] Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.
- [18] Wang, Z., & Culotta, A. (2021). Robustness to spurious correlations in text classification via automatically generated counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16), 14024-14031.

- [19] Wu, D., Wen, L., Chen, C., & Shi, Z. (2024, February). A novel counterfactual data augmentation method for aspect-based sentiment analysis. In Asian Conference on Machine Learning (pp. 1479-1493). PMLR.
- [20] Zhang, L., Zhang, X., Hu, L., Song, D., & Nie, L. (2025, July). Dually Self-Improved Counterfactual Data Augmentation Using Large Language Model. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5216-5227).
- [21] Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 107-117.
- [22] Chang, S., Zhang, Y., Yu, M., & Jaakkola, T. (2020, November). Invariant rationalization. In International Conference on Machine Learning (pp. 1448-1458). PMLR.
- [23] Yu, M., Chang, S., Zhang, Y., & Jaakkola, T. (2019, November). Rethinking cooperative rationalization: Introspective extraction and complement control. 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. 4094-4103).
- [24] Zhou, Y., & Zhu, Z. (2025). Fighting spurious correlations in text classification via a causal learning perspective. Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics. 4264–4274.
- [25] Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021). Counterfactual invariance to spurious correlations: Why and how to pass stress tests. arXiv preprint arXiv:2106.00545.
- [26] Weston, J., Collobert, R., Sinz, F., Bottou, L., & Vapnik, V. (2006, June). Inference with the universum. In Proceedings of the 23rd international conference on Machine learning (pp. 1009-1016).
- [27] Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. Advances in Neural Information Processing Systems, 22, 1410-1418.
- [28] Pushp, P. K., & Srivastava, M. M. (2017). Train once, test anywhere: Zero-shot learning for text classification. arXiv. <https://arxiv.org/abs/1712.05972>
- [29] Scheirer, W. J., Rocha, A., Micheals, R. J., & Boulton, T. E. (2011). Meta-recognition: The theory and practice of recognition score analysis. IEEE transactions on pattern analysis and machine intelligence, 33(8), 1689-1695.
- [30] Bendale, A., & Boulton, T. E. (2016). Towards open set deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1563-1572.
- [31] Qin, C., Chen, X., Li, J., & Zhu, H. (2026). BOLT: Benchmarking open-world learning for text classification. Findings of the Association for Computational Linguistics: ACL 2026, 13624-13658.
- [32] Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017, July). icarl: Incremental classifier and representation learning. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 5533-5542). IEEE.
- [33] Li, Z., & Hoi, S. C. H. (2017). Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12), 2935-2947.
- [34] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.
- [35] Xu, R., & Ding, K. (2025). Large language models for anomaly and out-of-distribution detection: A survey. Findings of the Association for Computational Linguistics: NAACL 2025, 6007-6027.
- [36] Kwon, S. M., Xu, A. S., Yaras, C., Balzano, L., & Qu, Q. (2026, May). Towards Understanding Out-of-Distribution Generalization for In-Context Learning via Low-Dimensional Subspaces. AI Stats Workshop on Causality in the Age of AI Scaling.

- [37] Qi, Z., Luo, H., Huang, X., Zhao, Z., Jiang, Y., Fan, X., ... & Glass, J. R. (2025, May). Quantifying generalization complexity for large language models. In *International Conference on Learning Representations* (Vol. 2025, pp. 97091-97113).
- [38] Sun, Y., Hu, S., Zhou, G., Zheng, K., Hajishirzi, H., Dziri, N., & Song, D. (2026). Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *Advances in Neural Information Processing Systems*, 38.
- [39] Suresh, P., Stanley, J., Joseph, S., Scimeca, L., & Bzdok, D. (2026). At the Edge of Understanding: Sparse Autoencoders Trace The Limits of Transformer Generalization. *arXiv preprint arXiv:2606.26396*.
- [40] Sheikh, T. U., Sinha, S., Sam, S., Stricker, D., & Afzal, M. Z. (2025, October). ZOD: Zero-Shot and Out-of-Distribution Detection Dataset for Document Images. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 7545-7555). IEEE.
- [41] Arora, U., Huang, W., & He, H. (2021, November). 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).
- [42] Baran, M., Baran, J., Wójcik, M., Zięba, M., & Gonczarek, A. (2023, July). Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 119-129).
- [43] Lake, B., & Baroni, M. (2018, July). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning* (pp. 2873-2882). PMLR.
- [44] Kim, N., & Linzen, T. (2020, November). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 9087-9105).
- [45] Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., ... & Bousquet, O. (2019). Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.
- [46] Li, C., Cheng, K., Hao, W., Shao, T., & Zhang, J. (2026). Survey of compositional generalization methodologies in artificial intelligence. *CAAI Transactions on Intelligent Systems*, 21(4), 815-833.
- [47] Costache, A. V., Dragomir, D. N., Gheorghe, S. F., Poesina, E., Irofti, P., & Ionescu, R. T. (2026). MOSLD-Bench: Multilingual Open-Set Learning and Discovery Benchmark for Text Categorization. *arXiv preprint arXiv:2601.13437*.
- [48] Larson, S., Nagy, A., Desai, S., Desai, C., Lima, N., Yuan, Y., ... & Leach, K. (2026, August). Revising RVL-CDIP: Quantifying Errors and Test-Train Overlap. In *Proceedings of the 2026 ACM Symposium on Document Engineering* (pp. 1-10).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).