MOSLD-Bench: Multilingual Open-Set Learning and Discovery Benchmark for Text Categorization

Anonymous ACL submission

Abstract

Open-set learning and discovery (OSLD) is a challenging machine learning task in which samples from new (unknown) classes can appear at test time. It can be seen as a generalization of zero-shot learning, where the new classes are not known a priori, hence involving the active discovery of new classes. While zero-shot learning has been extensively studied in text classification, especially with the emergence of pre-trained language models, open-set learning and discovery is a comparatively new setup for the text domain. To this end, we introduce the first **m**ultilingual **o**pen-**s**et **l**earning and **d**iscovery (MOSLD) benchmark for text categorization by topic, comprising 960K data samples across 12 languages. To construct the benchmark, we (i) rearrange existing datasets and (ii) collect new data samples from the news domain. Moreover, we propose a novel framework for the OSLD task, which integrates multiple stages to continuously discover and learn new classes. We evaluate several language models, including our own, to obtain results that can be used as reference for future work. We release our benchmark at `https://anonymous.4open.science/r/MOSLD-Bench-EFD0`.

1 Introduction

In machine learning, training and test data are presumed to be independently and identically sampled from the same distribution (Vapnik, 1995). However, this scenario is often violated in practice. Real-world data can exhibit one or multiple types of distribution shifts, rendering training and test data dissimilar. Various types of shifts lead to various task formalizations (Costache et al., 2026), such as classification with background class (Dhamija et al., 2018; Hendrycks et al., 2019; Lee et al., 2018; Liu et al., 2020), zero-shot learning (Gera et al., 2022; Meng et al., 2020, 2022; Sanh et al., 2022; Yin et al., 2019; Zhang et al., 2024), and open-set learning (Chen et al., 2024; Geng et al., 2021; Scheirer et al., 2013).

Among these tasks, zero-shot learning (ZSL) is the most extensively studied, due to the spectacular advances obtained by pre-trained language models (Bommasani et al., 2021; Yang et al., 2025). ZSL is achieved either by adapting various related tasks in order to solve the classification problem (Yang et al., 2025; Yin et al., 2019; Zhou et al., 2024a), or by generating synthetic data and subsequently using a supervised method (Hu et al., 2017; Meng et al., 2022; Ye et al., 2022; Yu et al., 2023). While ZSL received significant attention in recent literature, it requires class labels to be known a priori, i.e. it does not cover the case where new classes that emerge during inference are completely unknown[1].

As highlighted by Costache et al. (2026), a more challenging task is *open-set learning and discovery* (OSLD) (Zheng et al., 2022), where new classes, unseen during training, gradually appear in the test data. The goal is to detect these new classes and later learn to recognize them, without degrading the performance on the existing classes. OSLD can be seen as a relaxed case of *open-set class incremental learning* (Xu et al., 2023). *Class incremental learning* (CIL) (Kim et al., 2022b; Masana et al., 2022; Zhou et al., 2024b) is a type of continual learning (Liu et al., 2023), where the problem is to learn a sequence of tasks, with each task adding new classes. Yet, a notable oversight in the realm of CIL is the inability to process the occurrence of unknown classes during inference. Open-set class incremental learning (Xu et al., 2023) merges class incremental learning and open-set recognition in a unified task formulation, where models must adapt to unknown classes, while access to already processed training data is restricted. In contrast, OSLD does not impose restricted access to training data, rendering catastrophic forgetting (Kirkpatrick

[1]Please see the distinction between Unknown Known Classes (UKCs) and Unknown Unknown Classes (UUCs) made by Geng et al. (2021).

Language	Source	Train		Val		Test 1		Test 2		Test 3		Total	
		#classes	#samples	#classes	#samples	#classes	#samples	#classes	#samples	#classes	#samples	#classes	#samples
Arabic	Ultimate Arabic News	4	98,350	4	14,341	6	18,341	8	24,953	10	32,664	10	188,649
Bengali	L3Cube	4	14,402	4	3,602	6	12,164	8	17,916	10	23,837	10	71,921
Chinese	THUCNews	4	22,367	4	7,998	7	13,555	10	25,174	14	52,346	14	121,440
English	DBpedia	4	40,000	4	4,000	7	21,000	10	30,000	14	42,000	14	137,000
French	Collected	4	42,534	4	5,962	6	11,255	8	17,998	10	21,388	10	99,137
Hungarian	Collected	4	10,323	4	1,141	6	7,527	8	12,250	10	16,943	10	48,184
Italian	Collected	4	10,238	4	1,809	6	5,936	8	7,753	10	9,730	10	35,466
Japanese	Collected	4	11,371	4	2,009	6	6,000	8	8,000	10	10,000	10	37,380
Romanian	MOROCO+collected	4	32,157	4	19,614	6	8,577	8	12,408	10	17,598	10	90,354
Russian	Rus News	4	11,248	4	2,813	6	8,204	8	11,673	10	15,781	10	49,719
Spanish	Collected	4	20,080	4	3,546	6	6,000	8	8,000	10	10,000	10	47,626
Turkish	Kemik	4	9,470	4	2,368	7	6,128	10	7,421	13	8,175	13	33,562
Total	Existing + collected	-	322,540	-	69,203	-	124,687	-	183,546	-	260,462	-	960,438

Table 1: Number of classes and number of samples for each language and each official subset in MOSLD-Bench.

et al., 2017) easy to address. Nevertheless, OSLD is arguably one the most difficult versions of text classification under class distribution shift (Costache et al., 2026), because the system must actively learn to identify new patterns in the data and adapt to them without any supervision. OSLD has practical motivation in text categorization, where new topics can emerge over time and systems need to handle them automatically. OSLD can also be seen as an extension of open-set learning (OSL), where samples from unknown classes have to be classified as such, in addition to being detected. While OSL has been explored in the text domain (Chen et al., 2023, 2024; Kim et al., 2022a; Walkowiak et al., 2020, 2019a,b), the more challenging OSLD problem remains largely unexplored in NLP literature (Costache et al., 2026).

To this end, we introduce the first **m**ultilingual **o**pen-**s**et **l**earning and **d**iscovery (MOSLD) benchmark for text categorization by topic. We first construct a comprehensive multilingual dataset via (i) rearranging existing datasets (such that some classes are not represented in the training set) and (ii) collecting new data samples from the news domain (for languages where existing resources are scarce). We further benchmark several small (e.g. BERT (Devlin et al., 2019)) and large (e.g. GPT-4o (Hurst et al., 2024)) language models on MOSLD-Bench. Since the task was not previously explored, we propose a novel OSLD framework that integrates multiple stages (e.g. outlier detection, k-means clustering, TFIDF-based keyword extraction, BERT retraining with pseudo-labeling) to continuously discover and learn new classes.

In summary, our contribution is twofold:

- We introduce a novel multilingual dataset for open-set learning and discovery, which is obtained via new data gathering and existing dataset restructuring.
- We propose a new framework designed to address all the challenges of OSLD via a multi-stage processing pipeline based on keyword extraction, clustering, pseudo-labeling and model retraining.

2 Dataset

Task definition. Let $\mathcal{D} = \{(x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$ be a training set, where x is data sample from the space $\mathcal{X}$, and y is a label from the space $\mathcal{Y}$. The goal of OSLD is to predict the labels for some test set $\mathcal{T}_i = \{(x, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}_i^+\}$, where i represents the test set index, which increases over time, and $\mathcal{Y}_i^+$ contains class labels that are not in $\mathcal{Y}$, i.e. $\mathcal{Y} \subset \mathcal{Y}_i^+$. At any timestep i, the unlabeled test sets $\mathcal{T}_1, ..., \mathcal{T}_i$ and the original training set $\mathcal{D}$ can be used to obtain the model $h_i : \mathcal{X} \to \mathcal{Y}_i^+$, which is supposed to correctly identify samples belonging to $\mathcal{Y}$, to discover classes from $\mathcal{Y}_i^+ \setminus \mathcal{Y}$, and recognize samples from the discovered classes.

Overview. Following the task definition above, we construct a multilingual dataset for OSLD, such that for each language, only a subset of the total number of classes is included in the training data, while the remaining classes are gradually inserted during testing. More precisely, for each language, we construct three test sets, such that $\mathcal{Y}_i^+ \subset \mathcal{Y}_{i+1}^+, \forall i \in \{1, 2\}$. The dataset contains 960,438 samples from 12 typologically diverse languages, covering multiple language families and diverse writing systems (see Table 1).

Data collection. The corpus is compiled from two complementary sources. For high-resource languages, we integrate publicly available datasets: Ultimate Arabic News Dataset (Al-Dulaimi, 2022) for Arabic, L3Cube-IndicNews (Mirashi et al., 2023) for Bengali, DBpedia Ontology Dataset (Zhang et al., 2015) for English, THUCNews (Sun et al., 2016) for Chinese, Turkish News Kemik (Yıldırım and Yıldız, 2018) for Turkish, and Rus News Dataset (Kuznetsov, 2024) for Russian. For Romanian, we extended the MOROCO dataset (Butnaru and Ionescu, 2019) (originally containing 6 categories) by collecting additional news articles for 4 new categories, resulting in a total of 10 categories. For languages lacking suitable public resources (French, Italian, Japanese, Hungarian, and Spanish), we collect news articles directly from local news portals. The collected data was subsequently anonymized to remove personally identifiable information. Category labels were derived from the existing editorial categories used by each news website, requiring no manual annotation. In the end, each language subset contains between 10 and 14 categories covering common news domains, such as politics, sport, economy, technology, culture, health, etc.

Dataset organization. To enable standardized evaluation, we apply a uniform splitting protocol across all languages. For each language, we designate 4 classes as "known" from the beginning (belonging to $\mathcal{Y}$), splitting their samples into training, validation, and test. The remaining classes are treated as "unknown" (belonging to $\mathcal{Y}_i^+ \setminus \mathcal{Y}$). These unknown classes are progressively introduced across three evaluation stages. Test $\mathcal{T}_1$ contains test samples from the 4 baseline classes plus samples from 2-3 unknown classes. Test $\mathcal{T}_2$ contains test samples from the baseline classes, samples from the classes introduced in $\mathcal{T}_1$, plus samples from another 2-3 new classes. Test $\mathcal{T}_3$ follows the same pattern, incorporating the rest of unknown classes. Once a class is introduced, its samples remain present in all subsequent test sets. This structure enables evaluation of both the model's ability to maintain accuracy on known classes and its capacity to detect novel, previously unseen categories. In Table 1, we provide detailed information about the number of classes and the number of samples per language and per subset, respectively.

Evaluation procedure. We evaluate models on two complementary objectives: maintaining performance on previously learned classes, and performing well on novel categories. We report accuracy rates and macro-averaged F1 scores for three samples groups: (1) **overall** – includes all test samples in $\mathcal{T}_i$, (2) **known** – includes samples from classes seen during training, (3) **unknown** – includes samples from classes introduced in the current test set $\mathcal{T}_i$, for all $i \in \{1, 2, 3\}$. We track accuracy on each group across successive stages, measuring how performance on earlier classes degrades as new categories are learned. Since discovered classes are learned in an unsupervised manner, without access to ground-truth labels, we determine the correspondence between discovered and ground-truth labels using Hungarian matching. More details about the matching procedure are given in Appendix A.1.

3 Methods

Baseline based on known-class supervision. To establish a performance lower bound, we fine-tune language-specific pre-trained BERT encoders. The specific version of BERT for each language is reported in Table 3 from Appendix A.2. The models are optimized via standard supervised learning (cross-entropy loss) only on the initial set of known classes $\mathcal{Y}$. The weights stay frozen after the initial training phase, while the models are evaluated sequentially on all three test sets. This setting corresponds to a classifier that does not adapt to distribution and label-space shift. The corresponding performance levels serve as a reference point for OSLD methods that learn continuously.

Proposed OSLD methods. We propose two alternative OSLD methods that follow a joint processing pipeline comprising multiple stages: (1) outlier sample detection, (2) outlier data clustering, (3) class-specific keyword extraction, (4) model retraining. These steps are sequentially applied on each test set $\mathcal{T}_i$. We start from the language-specific BERT models fine-tuned on the training set (comprising only known classes). In step (1), we identify samples not belonging to any learned (known) category using energy-based detection. The energy score of input x is computed as:

$$E(x) = -\log\left(\sum_{j=1}^{N} \exp\left(h^j(x)\right)\right), \quad (1)$$

where $h^j(x)$ denotes the j-th logit output of the classifier h and N is the number of currently known classes. Intuitively, samples from known classes yield lower energy scores due to higher confidence predictions, while outliers produce higher energy values. Therefore, samples with a high energy (top 15%) are labeled as outliers, while the remaining samples are assigned to one of the known classes. In step (2), we cluster outliers using k-means on [CLS] embeddings provided by the language model. The optimal k is determined by max-

Language	Test 1 Baseline Acc / F1	Test 1 V1 Acc / F1	Test 1 V2 Acc / F1	Test 2 Baseline Acc / F1	Test 2 V1 Acc / F1	Test 2 V2 Acc / F1	Test 3 Baseline Acc / F1	Test 3 V1 Acc / F1	Test 3 V2 Acc / F1
Arabic	0.517 / 0.371	**0.885 / 0.881**	0.867 / 0.867	0.370 / 0.226	0.586 / 0.616	**0.607 / 0.631**	0.285 / 0.154	**0.605 / 0.568**	0.581 / 0.529
Bengali	0.448 / 0.301	0.497 / 0.527	**0.508 / 0.539**	0.302 / 0.160	**0.350 / 0.368**	0.333 / 0.359	0.226 / 0.105	0.266 / 0.272	**0.284 / 0.291**
Chinese	0.396 / 0.233	0.725 / 0.742	**0.752 / 0.763**	0.207 / 0.082	0.654 / **0.642**	**0.655** / 0.641	0.099 / 0.019	0.436 / 0.434	**0.475 / 0.458**
English	0.570 / 0.467	0.723 / 0.692	**0.746 / 0.713**	0.399 / 0.301	0.512 / 0.482	**0.546 / 0.511**	0.285 / 0.165	0.430 / 0.362	**0.442 / 0.378**
French	0.340 / 0.178	0.892 / 0.877	**0.896 / 0.881**	0.213 / 0.079	**0.575 / 0.623**	0.572 / 0.572	0.179 / 0.057	0.407 / 0.442	**0.454 / 0.472**
Hungarian	0.512 / 0.364	**0.680 / 0.715**	0.668 / 0.699	0.314 / 0.160	**0.647 / 0.666**	0.630 / 0.658	0.227 / 0.093	0.525 / 0.552	**0.537 / 0.559**
Italian	0.632 / 0.529	**0.707 / 0.693**	0.698 / 0.651	0.482 / 0.351	0.573 / 0.533	**0.583 / 0.551**	0.386 / 0.243	**0.449** / 0.414	0.445 / **0.416**
Japanese	0.622 / 0.508	**0.625 / 0.595**	0.624 / 0.583	**0.470** / 0.332	0.446 / **0.438**	0.443 / 0.434	0.374 / 0.232	**0.439 / 0.382**	0.429 / 0.376
Romanian	**0.452** / 0.370	0.448 / **0.531**	0.448 / 0.530	0.312 / 0.203	0.365 / 0.436	**0.370 / 0.442**	0.219 / 0.137	**0.230 / 0.308**	0.228 / 0.306
Russian	0.526 / 0.383	0.625 / 0.649	**0.639 / 0.660**	0.366 / 0.210	**0.536 / 0.531**	0.523 / 0.518	0.274 / 0.128	**0.380 / 0.391**	0.375 / 0.386
Spanish	**0.640** / 0.522	0.634 / **0.595**	0.632 / 0.578	0.480 / 0.335	0.538 / 0.487	**0.541 / 0.496**	0.385 / 0.239	0.431 / 0.362	**0.435 / 0.377**
Turkish	0.518 / **0.434**	**0.534** / 0.432	0.529 / 0.430	**0.429 / 0.328**	0.411 / 0.264	0.413 / 0.267	**0.393 / 0.289**	0.365 / 0.202	0.361 / 0.202

Table 2: **Overall** accuracy and F1 scores across 12 languages for each evaluation stage in MOSLD-Bench. The best score for each language and each test set is highlighted in bold.

imizing the silhouette coefficient. The resulting k represents the number of discovered classes. In step (3), we extract a list of representative keywords for each cluster, which can be further used to match the respective cluster with one of the ground-truth classes. To extract keywords, we merge all the samples in each cluster into a single document, then apply the TFIDF scheme over the resulting documents. This promotes words that appear in only one of the clusters. We select the top 10 keywords for each cluster. Each keyword list is passed through BERT and the corresponding `[CLS]` embedding becomes a centroid for the corresponding cluster. In step (4), we sort the samples in each cluster based on the cosine similarity with their centroid. We keep 40% of samples closest to each centroid for model retraining. At this stage, we implement two alternative approaches. In the first approach (V1), the set of classes is expanded with the discovered clusters and the model is retrained with standard cross-entropy. In the second approach (V2), we augment cross-entropy with a contrastive term that pulls sample embeddings towards their centroid, grounding representations in semantic information from clustering:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{CL}}. \tag{2}$$

$\mathcal{L}_{\text{CE}}$ is the cross-entropy and $\mathcal{L}_{\text{CL}}$ is defined as:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \log \left(\frac{\exp(\text{sim}(\mathbf{e}_i, \mathbf{c}_{\hat{y}_i})/\tau)}{\sum_{j=1}^{k} \exp(\text{sim}(\mathbf{e}_i, \mathbf{c}_j)/\tau)} \right), \tag{3}$$

where $\mathcal{U}$ denotes the set of samples from discovered classes, $\mathbf{e}_i$ is the `[CLS]` embedding of sample x_i, $\mathbf{c}_{\hat{y}_i}$ is the centroid of the cluster $\hat{y}_i$ that includes sample x_i, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is a temperature parameter.

4 Experiments

Hyperparameter tuning. To reproduce results, we provide details about hyperparameter choices in Appendix A.3.

Baseline vs. OSLD BERT models. In Table 2, we present the accuracy and F1 scores for all classes (**overall**), across all languages and evaluation stages. The baseline model, which remains frozen after initial training on known classes, shows consistent performance degradation as new classes are introduced. This is expected, since the baseline cannot recognize samples from unknown categories, assigning them to one of the known classes. In contrast, the proposed OSLD methods (V1 and V2) demonstrate the ability to discover and learn new classes, generally achieving substantially higher scores than the baseline. Comparing V1 and V2, we observe that integrating the contrastive loss (V2) provides improvements on several languages, particularly in early stages (T1 and T2). We present results for **known** and **unknown** classes in Tables 4 and 5, discussing them in Appendix A.4.

Small vs. large language models. We discuss this comparative study in Appendix A.5.

5 Conclusion

We introduced a multilingual dataset to benchmark methods for open-set learning and discovery. To construct the dataset, we adopted two approaches: (i) reorganizing existing datasets and (ii) collecting new data from the web. We evaluated a set of baseline methods, creating reference points for future research on the topic. Since OSLD is a new task in the NLP domain, we also proposed a method specifically designed to solve the OSLD task.

In future work, we aim to evaluate additional OSLD methods on MOSLD-Bench.

6 Limitations

Our evaluation includes a limited number of baseline models, reflecting the scarcity of methods explicitly designed for open-set learning and discovery. According to Costache et al. (2026), OSLD has not been explored in the NLP domain. We aim to address this limitation by closely tracking future work on OSLD and include newly developed models.

In addition, the chosen large language model, GPT-4o, is evaluated in a reduced setup, comprising two languages and only one test stage. This is due to the high cost of the API-based inference with GPT-4o.

References

Ahmed Hashim Al-Dulaimi. 2022. Ultimate Arabic News Dataset. Mendeley Data, V1.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Andrei M. Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 688–698.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Proceedings of Practical ML for Developing Countries Workshop at ICLR 2020*.

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

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

Adriana Valentina Costache, Silviu Florin Gheorghe, Eduard Gabriel Poesina, Paul Irofti, and Radu Tudor Ionescu. 2026. A survey of text classification under class distribution shift. In *Proceedngs of 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

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

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

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

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

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. 2022b. A theoretical study on solving continual learning. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 5065–5079.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Maxim Kuznetsov. 2024. Russian News Classifier Dataset. Hugging Face Datasets.

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

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mihai Masala and Stefan Ruseti. 2020. RoBERT – A Romanian BERT Model. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 5879–5887.

Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. 2022. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533.

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

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

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2023. L3Cube-IndicNews: News-based Short Text and Long Document Classification Datasets in Indic Languages. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 442–449.

Dávid Márk Nemeskey. 2021. Introducing huBERT. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY)*, pages 3–14, Szeged, Hungary.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, and 22 others. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sagor Sarker. 2020. BanglaBERT: Bengali Mask Language Model for Bengali Language Understanding.

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

Stefan Schweter. 2020a. BERTurk – BERT models for Turkish. Zenodo.

Stefan Schweter. 2020b. Europeana BERT: Pretrained Transformer Models for Historical Text. Zenodo.

Maosong Sun, Jingyang Li, Zhipeng Guo, Zhao Yu, Yabin Zheng, Xiance Si, and Zhiyuan Liu. 2016. THUCTC: An Efficient Chinese Text Classifier. `http://thuctc.thunlp.org/`.

Tohoku NLP. 2023. bert-base-japanese-v3: A Japanese BERT Model. `https://huggingface.co/tohoku-nlp/bert-base-japanese-v3`. Accessed: 2026-01-05.

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

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

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

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

Jiawen Xu, Claas Grohnfeldt, and Odej Kao. 2023. OpenIncrement: A unified framework for open set recognition and deep class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3303–3311.

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

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

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

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

Savaş Yıldırım and Tuğba Yıldız. 2018. A Comparative Analysis of Text Classification for Turkish Language. *Pamukkale University Journal of Engineering Sciences*, 24(5):879–886.

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

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

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

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

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2024b. Class-Incremental Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, Tatiana Shavrina, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 507–524.

Language	Model
Arabic	AraBERT (Antoun et al., 2020)
Bengali	BanglaBERT (Sarker, 2020)
Chinese	Chinese BERT (Devlin et al., 2019)
English	BERT (Devlin et al., 2019)
French	French BERT (Schweter, 2020b)
Hungarian	huBERT (Nemeskey, 2021)
Italian	Italian BERT (Schweter, 2020b)
Japanese	TohokuBERT (Tohoku NLP, 2023)
Romanian	RoBERT(Masala and Ruseti, 2020)
Russian	ruBERT (Zmitrovich et al., 2024)
Spanish	Spanish BERT (Cañete et al., 2020)
Turkish	BERTurk (Schweter, 2020a)

Table 3: All models use the base architecture, adapted for uncased text processing.

A Appendix

A.1 Discovered and Ground-Truth Class Matching

To determine the correspondence between discovered and ground-truth labels, we compute the semantic similarity between discovered and ground-truth class names. For each discovered class, we compute an embedding representation from its characteristic keywords (a list of representative words returned by the OSLD method). We then measure cosine similarity between embeddings of discovered and ground-truth class names. The embeddings are obtained from language-specific pretrained transformer models (see Table 3). The matching is performed exclusively during the evaluation stage to assess discovery quality, while ground-truth labels remain unused during training or inference. The optimal one-to-one assignment between discovered and ground-truth classes is obtained using Hungarian matching, enabling a fully automated evaluation, without any manual intervention.

A.2 Language-Specific BERT Models

In Table 3, we specify the BERT version used as the starting point for the fine-tuning process, for each language. Preliminary experiments with multilingual BERT (mBERT) (Devlin et al., 2019) led to

Language	Test 1			Test 2			Test 3		
	Baseline	V1	V2	Baseline	V1	V2	Baseline	V1	V2
	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1
Arabic	**0.988 / 0.988**	0.955 / 0.638	0.961 / 0.639	**0.986 / 0.986**	0.629 / 0.556	0.661 / 0.572	**0.985 / 0.985**	0.548 / 0.496	0.520 / 0.458
Bengali	**0.908 / 0.908**	0.810 / 0.548	0.813 / 0.554	**0.902 / 0.901**	0.439 / 0.389	0.416 / 0.375	**0.900 / 0.900**	0.302 / 0.302	0.308 / 0.289
Chinese	**0.989 / 0.989**	0.954 / 0.646	0.976 / 0.654	**0.989 / 0.989**	0.686 / 0.573	0.678 / 0.572	**0.989 / 0.989**	0.586 / 0.534	0.596 / 0.539
English	**0.998 / 0.998**	0.931 / 0.639	0.959 / 0.651	**0.998 / 0.998**	0.647 / 0.501	0.689 / 0.528	**0.999 / 0.999**	0.442 / 0.370	0.459 / 0.384
French	**0.963 / 0.963**	0.948 / 0.762	0.944 / 0.763	**0.966 / 0.966**	0.715 / 0.616	0.706 / 0.706	**0.967 / 0.967**	0.451 / 0.460	0.456 / 0.462
Hungarian	**0.972 / 0.972**	0.942 / 0.636	0.947 / 0.638	**0.971 / 0.971**	0.630 / 0.536	0.663 / 0.548	**0.972 / 0.972**	0.451 / 0.493	0.495 / 0.494
Italian	**0.938 / 0.938**	0.892 / 0.610	0.911 / 0.613	**0.935 / 0.935**	0.607 / 0.470	0.627 / 0.487	**0.939 / 0.938**	0.481 / 0.390	0.468 / 0.468
Japanese	**0.933 / 0.933**	0.908 / 0.619	0.918 / 0.619	**0.941 / 0.941**	0.578 / 0.438	0.576 / 0.437	**0.936 / 0.936**	0.431 / 0.347	0.417 / 0.339
Romanian	**0.970 / 0.970**	0.960 / 0.645	0.962 / 0.645	**0.969 / 0.969**	0.443 / 0.394	0.442 / 0.393	**0.967 / 0.967**	0.333 / 0.319	0.331 / 0.317
Russian	**0.921 / 0.921**	0.876 / 0.722	0.892 / 0.729	**0.913 / 0.912**	0.563 / 0.478	0.550 / 0.464	**0.922 / 0.922**	0.397 / 0.391	0.394 / 0.386
Spanish	**0.961 / 0.961**	0.934 / 0.632	0.943 / 0.633	**0.960 / 0.960**	0.624 / 0.452	0.623 / 0.457	**0.962 / 0.962**	0.469 / 0.352	0.473 / 0.364
Turkish	**0.805 / 0.799**	0.774 / 0.512	0.766 / 0.509	**0.806 / 0.801**	0.491 / 0.309	0.492 / 0.313	**0.814 / 0.809**	0.393 / 0.216	0.387 / 0.215

Table 4: Accuracy and F1 scores computed of **known** classes across 12 languages for each evaluation stage in MOSLD-Bench. The best score for each language and each test set is highlighted in bold.

Language	Test 1			Test 2			Test 3		
	Baseline	V1	V2	Baseline	V1	V2	Baseline	V1	V2
	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1	Acc / F1
Arabic	0 / 0	0.477 / 0.095	**0.759 / 0.275**	0 / 0	**0.586 / 0.616**	0.471 / 0.090	0 / 0	**0.761 / 0.117**	0.756 / **0.117**
Bengali	0 / 0	0.192 / 0.080	**0.205 / 0.086**	0 / 0	**0.161 / 0.052**	0.156 / 0.050	0 / 0	0.155 / 0.049	**0.210 / 0.060**
Chinese	0 / 0	0.572 / 0.207	**0.602 / 0.218**	0 / 0	0.619 / 0.174	**0.629 / 0.175**	0 / 0	0.296 / 0.072	**0.363 / 0.090**
English	0 / 0	0.451 / 0.161	**0.463 / 0.166**	0 / 0	**0.280 / 0.095**	0.212 / 0.070	0 / 0	0.384 / 0.091	**0.397 / 0.103**
French	0 / 0	0.861 / 0.306	**0.870 / 0.308**	0 / 0	0.342 / **0.103**	**0.350** / 0.090	0 / 0	0.168 / 0.036	**0.443 / 0.089**
Hungarian	0 / 0	**0.389 / 0.115**	0.357 / 0.104	0 / 0	**0.675 / 0.197**	0.577 / 0.143	0 / 0	0.626 / **0.164**	**0.644** / 0.151
Italian	0 / 0	**0.323 / 0.123**	0.259 / 0.070	0 / 0	**0.462 / 0.113**	0.438 / 0.107	0 / 0	0.321 / 0.070	**0.353 / 0.403**
Japanese	0 / 0	**0.058 / 0.024**	0.037 / 0.015	0 / 0	**0.051 / 0.013**	0.043 / 0.011	0 / 0	0.468 / 0.069	**0.480 / 0.072**
Romanian	0 / 0	**0.003 / 0.001**	0.002 / **0.001**	0 / 0	0.192 / **0.092**	**0.209** / 0.085	0 / 0	0 / 0	0 / 0
Russian	0 / 0	**0.304 / 0.088**	0.282 / 0.083	0 / 0	**0.471** / 0.105	0.459 / **0.107**	0 / 0	**0.331** / 0.074	0.320 / **0.075**
Spanish	0 / 0	**0.034 / 0.013**	0.008 / 0.003	0 / 0	0.278 / **0.077**	**0.294** / 0.076	0 / 0	0.278 / 0.056	**0.284 / 0.057**
Turkish	0 / 0	**0.102** / 0.070	0.100 / **0.076**	0 / 0	0.034 / **0.015**	**0.037 / 0.015**	0 / 0	0.084 / 0.031	**0.104 / 0.035**

Table 5: Accuracy and F1 scores computed of **unknown** classes across 12 languages for each evaluation stage in MOSLD-Bench. The best score for each language and each test set is highlighted in bold.

significantly worse results in several languages. To address this problem, we selected language-specific BERT models. The same models are used for both the naive approach (which ignores new classes) and the proposed OSLD approaches, ensuring a fair comparison between methods.

A.3 Hyperparameter Tuning

We use consistent hyperparameters across all languages to ensure a fair comparison. All models are fine-tuned using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2{\cdot}10^{-5}$, weight decay of 0.01, and linear warmup over the first 100 steps. We train for 5 epochs with a batch size of 16. During clustering, we search for the optimal k between 2 and 8. We set $\lambda = 0.3$ in Eq. (2), and $\tau = 0.07$ in Eq. (3). We use the default values for all other hyperparameters.

A.4 Performance on Known vs. Unknown Classes

In Tables 4 and 5, we provide a detailed analysis of model performance, separating **known** classes (seen during training) from **unknown** classes (discovered during testing).

Known classes. The baseline model achieves high accuracy on known classes across all languages and evaluation stages, typically exceeding 90%. This confirms that the frozen baseline effectively retains its initial knowledge, which is expected when no model updates occur. Both V1 and V2 maintain competitive performance on known classes on the first test stage ($\mathcal{T}_1$), with only minor drops with respect to the baseline. For subsequent performance stages, performance gaps become more evident.

Unknown classes. The baseline trivially scores zero on unknown classes, as it lacks the capacity to predict categories outside its training distribution. In contrast, V1 and V2 demonstrate the ability to discover and classify novel categories, with varying degrees of success across languages. The OSLD methods exhibit strong performance on Arabic, with V2 achieving 75.9% accuracy on $\mathcal{T}_1$ and maintaining 75.6% on $\mathcal{T}_3$. For French, both V1 and V2 models achieve impressive performance on $\mathcal{T}_1$, but this level of performance is not maintained on $\mathcal{T}_2$ and $\mathcal{T}_3$.

(a) Prompt for label generation
You must behave as an Open-Set Classification Model which discovers the class of the presented texts. Each text has only one category. You will be provided the texts, numbered by their id. The known text classes are: <LIST_OF_CLASS_NAMES>.
Your task is to discover how many new classes you can identify and name them alongside the existing ones. You must respond only with an array of all the existing classes. Do not use third party tools or python code to identify the names of the classes. Use only predictions based on what you read. Output only the classes.
(b) Prompt for sample classification
You need to behave as a text classification model.
You need to use the following classes in order to classify the texts: <LIST_OF_CLASS_NAMES>
For each text example, output ONLY the class id (a single number from 0–<NUM_OF_CLASSES>). Nothing else.

Table 6: Prompts used for GPT-4o: (a) open-set discovery and generation of class labels, and (b) classification of text samples. Text files containing training and test samples were attached to both prompts.

Language	Test 1			
	Baseline	V1	V2	GPT-4o
French	0.340 / 0.178	0.892 / 0.877	**0.896 / 0.881**	0.713 / 0.630
Turkish	0.518 / **0.434**	**0.534** / 0.432	0.529 / 0.430	0.454 / 0.362

Table 7: **Overall** accuracy and F1 scores for BERT-based methods (baseline, V1 and V2) versus GPT-4o. Results are reported for the $\mathcal{T}_1$ test set. The best score for each language is highlighted in bold.

Known vs. unknown classes. Comparing the results in Tables 4 and 5, we observe that all models exhibit considerable performance degradation when going from known classes to unknown classes. This observation highlights the difficulty of discovering and learning new classes without supervision. Overall, the empirical results confirm that MOSLD-Bench is a challenging benchmark, and the OSLD task remains open for future work.

A.5 Small vs. Large Language Models

We further compare the BERT-based models with a large language model (Hurst et al., 2024) evaluated in the OSLD setting. More precisely, GPT-4o is provided with the test data, and asked to automatically discover new classes and label test instances. This is comparatively harder than a typical zero-shot setting, where the model is given all class names. We provide the set of prompts used for GPT-4o in Table 6.

The comparison between BERT-based and GPT-4o language models is presented in Table 7. Due to the high pricing of LLM inference, this analysis is restricted to the first evaluation stage ($\mathcal{T}_1$), and only two languages (French and Turkish).

While GPT-4o demonstrates strong OSLD capabilities, we observe a tendency towards leveraging data leakage. For Turkish, the model immediately proposes nearly the full set of original labels, suggesting reliance on prior knowledge acquired during pre-training, rather than actual discovery from the test distribution. A similar behavior is observed for French, where the model tends to predict a large number of distinct categories, indicating a bias toward over-generation that may be sourced from knowledge acquired during large-scale pre-training. Quantitatively, the proposed OSLD frameworks (V1 and V2) outperform the large language model, while operating under substantially lower computational cost. These findings highlight a trade-off between the expressive power of LLMs and the methodological alignment of smaller task-specific models.