Improved Out-of-Scope Intent Classification with Dual Encoding and Threshold-based Re-Classification

Hossam M. Zawbaa[1, 2], **Wael Rashwan**[1, 2, 3], **Sourav Dutta**[4], **Haytham Assem**[5]
[1]Technological University Dublin, Ireland.
[2] Ireland's Centre for Applied AI (CeADAR), Ireland
[3] 3S Group, Faculty of Business, TU DUBLIN, Ireland
[4]Huawei Research Centre, Ireland
[5]Amazon Alexa AI, United Kingdom
hossam.zawbaa@gmail.com, wael.rashwan@tudublin.ie, sourav.dutta2@huawei.com,
hithsala@amazon.co.uk

Abstract

Detecting out-of-scope user utterances is essential for task-oriented dialogues and intent classification. Current methodologies face difficulties with the unpredictable distribution of outliers and often rely on assumptions about data distributions. We present the Dual Encoder for Threshold-Based Re-Classification (DETER) to address these challenges. This end-to-end framework efficiently detects out-of-scope intents without requiring assumptions on data distributions or additional post-processing steps. The core of DETER utilizes dual text encoders, the Universal Sentence Encoder (USE) and the Transformer-based Denoising AutoEncoder (TSDAE), to generate user utterance embeddings, which are classified through a branched neural architecture. Further, DETER generates synthetic outliers using self-supervision and incorporates out-of-scope phrases from open-domain datasets. This approach ensures a comprehensive training set for out-of-scope detection. Additionally, a threshold-based re-classification mechanism refines the model's initial predictions. Evaluations on the CLINC-150, Stackoverflow, and Banking77 datasets demonstrate DETER's efficacy. Our model outperforms previous benchmarks, achieving an increase of up to 13% and 5% in F1 score for known and unknown intents on CLINC-150 and Stackoverflow, and 16% for known and 24% for unknown intents on Banking77. The source code has been released at https://github.com/Hossam-Mohammed-tech/Intent_Classification_OOS.

Keywords: Out-of-Scope Detection, Intent Classification, Dual Encoder, USE, TSDAE

1. Introduction

The rise of Generative AI models such as ChatGPT underscores the prominence of conversational frameworks as the primary interactive interface for end-user assistance. A crucial goal within these systems is to precisely discern the user's intent behind their utterances, a fundamental component for formulating appropriate responses (Coucke et al., 2018). However, the challenge lies in acquiring comprehensive training data that covers the diverse spectrum of potential user intents.

A competent intent classifier should identify the predefined intents and recognize *out-of-scope* (OOS) data. OOS data are inputs unrelated to the trained classes, representing queries that do not fit into predefined categories or intents (Cavalin et al., 2020). This is essential for truly understanding a user's objective.

Intent classifiers face challenges due to the requirement for domain-specific labelled datasets, particularly in low-data settings standard in commercial systems (Casanueva et al., 2020). Modern Language models like BERT aid in efficient intent classification by converting text into vector representations rich in semantic information. However, current strategies for Out-Of-Scope (OOS) intent detection often lack generalizability and refinement opportunities (Zhan et al., 2021). To address this, Zhan et al. (2021) introduced synthetic outlier construction using in-domain training data embeddings, but these methods are static and offer limited potential for refining model predictions.

In this paper, we introduce *Dual Encoder for Threshold-Based Re-Classification* (DETER), a method for efficient Out-Of-Scope (OOS) intent detection. DETER combines confidence thresholding and synthetic outlier generation to enhance intent classification performance (Zhan et al., 2021; Cer et al., 2018; Wang et al., 2021; Maas et al., 2013). Our neural network architecture utilizes dual sentence encoders: the Universal Sentence Encoder (USE) and Transformer-based Sequential Denoising Auto-Encoder (TSDAE) (Cer et al., 2018; Wang et al., 2021). We fine-tune the TSDAE model unsupervisedly using Clinc-150 data and incorporate synthetic outlier generation to enhance OOS intent detection (Zhan et al., 2021).

Summarizing our contributions, this research presents the innovative *Dual Encoder for Threshold-based Re-classification* (DETER) optimized for out-of-scope detection, offering a comprehensive solution devoid of stringent data distribution assumptions. The system adeptly captures an array of synaptic and semantic features by utilizing dual text encoding mechanisms, namely

LREC-COLING 2024, pages 8708–8718

USE and TSDAE, thereby refining user utterance representations and amplifying the efficiency of out-of-scope detection. A distinctive threshold-based re-classification mechanism further augments prediction reliability, ensuring a consistent and precise identification of out-of-scope intents. In addition, DETER employs a nuanced approach to synthetic outlier generation, merging self-supervision with open-domain datasets to cultivate a diversified training set. Empirical assessments on CLINC-150, Stackoverflow, and Banking77 datasets underscore its superior performance, with notable improvements in F1 scores relative to established benchmarks. Furthermore, DETER's streamlined architecture, with only 1.5 million trainable parameters, contrasts markedly with models like Zhan et al. (2021), boasting 125 million parameters, showcasing enhanced computational efficiency and scalability without compromising performance.

2. Related Literature

The choice of embeddings significantly impacts intent classification, with sentence-level embeddings gaining precedence over token-level ones in Natural Language Understanding (NLU). However, many specialized NLU models struggle outside core domains. Graph-based representations, which translate nodes to vector embeddings, hold promise for enhancing intent detection, particularly in Out-Of-Scope (OOS) classification. The potential of synthetic datasets, both artificial and real-world derived, is being explored to improve model adaptability across diverse data distributions. Techniques such as temperature scaling and strategic perturbations aid in distinguishing known from unknown samples. While Transformer models like BERT and RoBERTa have revolutionized NLU with their OOS resilience, optimization remains crucial. Additionally, emerging pre-trained dual encoders like USE and ConveRT further augment intent detection, with evidence suggesting their superiority over refinements such as BERT-Large.

The selection of embeddings holds critical importance in intent classification. TEXTOIR, a recent platform introduced for intent detection, provides tools for open intent detection and discovery tasks, offering a practical toolbox with expandable interfaces (Zhang et al., 2021a). For this study, we utilized open intent detection to distinguish n-class known intents and a one-class open intent, adhering to the methodologies and results published by TEXTOIR, using identical seeds to match intents in evaluating our DETER framework. OpenMax employs softmax loss and fits a Weibull distribution to classifier logits (Bendale and Boult, 2016), while MSP predicts known classes based on maximum softmax probabilities, excluding those with probabilities below 0.5 (Hendrycks and Gimpel, 2018). LOF utilizes density to pinpoint low-density outliers (Breunig et al., 2000), while DOC establishes multiple probability thresholds for known classes through Gaussian fitting (Shu et al., 2017). DeepUnk integrates margin loss with LOF for adept identification of unknown classes (Lin and Xu, 2019), and SEG utilizes a large margin Gaussian mixture loss for feature representation, offering a unique approach within deep learning techniques (Yan et al., 2020).

Diverging in methodology, ADB, a variant of DA-ADB, favours softmax loss over traditional distance-aware concepts (Zhang et al., 2023a, 2021b). Expanding on intent representations, the (K+1)-way method merges features from two in-domain intents (Zhan et al., 2021). MDF stands out on the outlier detection front by employing a one-class SVM post-Mahalanobis distance feature evaluation (Xu et al., 2021). ARPL innovates by maximizing variance between known-class samples and reciprocal point representations, setting its detection threshold at 0.5 (Chen et al., 2022). KNNCL captures attention by utilizing KNN, enhancing the learning of semantic features crucial for detecting out-of-scope intents (Zhou et al., 2022).

Training strategies, in particular, often become the defining factor for a model's success. A novel approach that has gained traction involves the infusion of synthetic outliers and the integration of OOS sentences from varied datasets during training (Zhan et al., 2021). Yet, specific strategies, like threshold introduction, may enhance OOS detection over BiLSTM baseline at the expense of in-domain performance (Hasani et al., 2022). Among these challenges, deep metric learning emerges as a novel approach in the intent detection space, fostering enhanced intent representation through techniques like triplet networks (Zhang et al., 2023b). Combined with advanced techniques like triplet loss and hard samples, these networks provide new precedents in discriminative intent representation.

3. Dual Encoder for Threshold-based Re-Classification (DETER)

This section describes the proposed DETER framework and discusses its internal workings.

Given *K* predefined intent classes (S_{known}) in a dialogue system, an intent detection model aims to predict an utterance's class, which may be one of the known intents or an out-of-scope intent. During the training phase, a set of *N-labeled utterances* (D_n^{tr}) is provided for training as shown below in Equations 1 and 2. This depicts a $K+1$ way classification problem at the test phase. Previous approaches commonly train a K-way classifier for the known intents and then attempt to find the outliers.

$$S_{known} = \{C_i\}_{i=1}^{K} \tag{1}$$

$$D_n^{\text{tr}} = \{(u_i, C_i)\}_{i=1}^{N} \mid C_i \in S_{known} \tag{2}$$

where C_i is the i^{th} intent class in the in-domain dataset, K is the total number of predefined intent classes, and u_i is the utterance.

On the other hand, as illustrated in Figure 1, DETER formalizes a (K + 1)-way classification task in the training phase by constructing out-of-scope examples through self-supervision and open-domain data without making data distribution assumptions. The dual encoder-based trained classifier can be easily used for inference without adaptation or post-processing. In the following subsections, we describe our proposed approach's details, including the dual encoders – Universal Sentence Encoder (USE) and Transformer-based Denoising AutoEncoder (TSDAE) – representation learning, construction of pseudo outliers, and training.

3.1. Universal Sentence Encoder (USE)

USE is a text embedding framework designed to transform text into high-dimensional vectors. Given a user utterance u, the USE embedding function, denoted by E_{USE}, maps u to a 512-dimensional vector in $R^{512}, i.e.,$E$_{\text{USE}} : u \to R^{512}$.

These embeddings have proven invaluable for many NLP tasks, including text classification, semantic similarity measurement, and clustering. The pre-trained USE is publicly accessible on Tensorflow-hub and is available in two forms, one trained with a Transformer encoder and the other with a Deep Averaging Network (DAN). The output is a 512-dimensional vector, while the input is the English text of varying length (Cer et al., 2018).

3.2. Transformer-based Denoising AutoEncoder (TSDAE)

TSDAE provides unsupervised training of the transformer architecture by introducing particular types of noise into the input text and subsequently reconstructing the vector representations (of the noisy input) to their original input values. During the training, TSDAE encodes broken sentences into fixed-sized vectors, and the decoder uses these sentence embeddings to rebuild the original sentence representations. The encoder's sentence embedding must accurately capture the semantics for higher-quality reconstruction.

We tune the TSDAE, built upon a RoBERTa transformer, in an unsupervised mode using the label-free CLINC-150 dataset, achieving domain adaptation (Wang et al., 2021). At inference, the encoder E_{TSDAE} transforms variable-length user utterance, u, into a fixed 768-dimensional vector Representation in R^{768}, $E_{\text{TSDAE}} : u \to R^{768}$.

Through noise-induced unsupervised tuning, TSDAE can effectively capture intricate sentence semantics, making it suitable for intent detection tasks and discerning user purposes from variable-length user utterances.

3.3. Representation Learning

To capture rich linguistic and contextual feature representation from user utterances, We employed the above two sentence embedding techniques: TSDAE and USE. Each user utterance, u, is independently processed through both the encoders, yielding a d-dimensional output vector of the TSDAE and USE, as shown in Equation 3.

$$h(u) = \begin{bmatrix} E_{\text{TSDAE}}(u) \\ E_{\text{USE}}(u) \end{bmatrix} \tag{3}$$

The dimensionality d of $h(u)$ amounts to 1280, which is attained by concatenating the embedding outputs from TSDAE (768 dimensions) and USE (512 dimensions). As a result, the user utterance transforms a vector $h(u)$ residing in R^{1280}. Post this representation learning; the training set is mapped to D_n^{tr} in the feature space as shown in equation 4.

$$D_n^{tr} = \left\{ (h_i, C_i) | \; h_i = \begin{bmatrix} E_{\text{TSDAE}}(u_i) \\ E_{\text{USE}}(u_i) \end{bmatrix}, \right. \\ \left. (u_i, c_i) \in D_n \right\}_{i=1}^{N} \tag{4}$$

3.4. Construction of Outliers

We implement the outlier construction method proposed by Zhan et al. (2021), which involves generating two training outliers: synthetic outliers via self-supervision and readily available open-domain outliers.

Synthetic Outliers We introduce hard outliers in the feature domain to enhance the model's generalization capability for out-of-scope (unknown) intent classification. These outliers are hypothesized to have representations analogous to established classes. Leveraging this hypothesis, we propose a self-supervised methodology to generate these hard outliers using the training set D_n^{tr} as delineated by Zhan et al. (2021).

Specifically, we construct synthetic outliers within the latent space by utilizing convex combinations of inlier features derived from disparate intent classes, as depicted in Equation 5.

$$h^{oos} = \theta \times h_\beta + (1 - \theta) \times h_\alpha \tag{5}$$

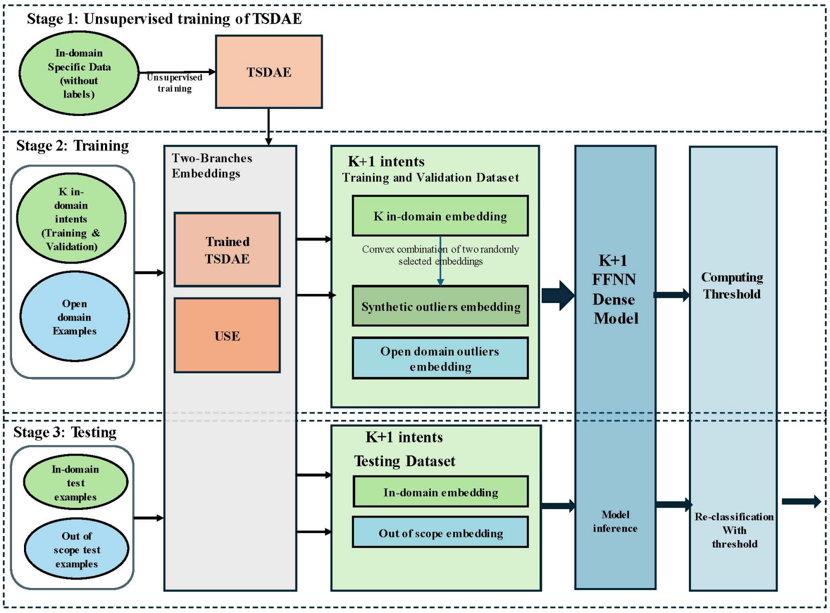

Figure 1: Overview of the proposed Dual Encoder for Threshold-based Re-Classification (DETER).

Where h_α and h_β are dual sentence encoder embeddings of two in-domain utterances randomly sampled from different intent classes in D_n^{tr}, and h^{oos} is the constructed synthetic outlier. θ is chosen arbitrarily from a uniform distribution $U(0; 1)$. To ensure diversity in synthetic Out-Of-Scope (OOS) training examples, we randomly sample θ within a specified range. This approach allows us to incorporate "hard," "medium," and "easy" examples for each iteration, thereby generating synthetic outliers that contribute to model training.

Open-Domain Outliers User input in practical dialogue systems can be arbitrary free-form sentences. Hence, learning architectures can easily use the available open-domain outliers to mimic real-world outliers and provide learning signals representing them during training. We thus model arbitrary outliers using the SQuaD 2.0 question-answering dataset (Rajpurkar et al., 2018) (Zhan et al., 2021). Combining the above outliers for training provides robustness to our framework for OOS detection. Later, we will study the effect of these outliers on DETER's performance.

3.5. DETER Training Architecture

The model architecture comprises two branches: one processing the Universal Sentence Encoder (USE) embedding of the input, while the other handles the Transformer-based TSDAE embedding. This refines the embeddings, aligning latent representations for similar intents. Each branch independently undergoes several dense and dropout layers to diminish the embedding dimension size. Subsequently, the condensed embeddings from both branches are concatenated before being forwarded to a classification head. Figure 2 illustrates the model's architecture. The entire system is optimized using cross-entropy loss against ground-truth intent labels from training data.

The aggregate count of trainable parameters in the proposed models amounts to 1,559,808 (1.5 M), in contrast to 124,645,632 (125 M) as observed in the K+1 method by Zhan et al. (2021).

3.6. Re-Classification Threshold

A key feature of our DETER framework involves enhancing model predictions through potential re-classification to enhance performance. In pursuit of this objective, we adopt the threshold-based approach proposed by Larson et al. (2019) for the softmax output of the classification model, as illustrated in Equation 6.

$$p(c|u) = softmax(Wh + b) \in R^{K+1} \tag{6}$$

where $W \in R^{(K+1)\times d}$ and $b \in R^{K+1}$ are the classifier's model parameters.

At inference time, we first take the prediction output of the model, i.e., the class C_j with the most significant value of $p(c = C_j|u)$. However, if the prediction confidence score of the model for the intent associated with an utterance ($I(u)$) is greater than a pre-computed threshold T, only then the model output is considered. Else, DETER tags the intent OOS. This is computed as in Equation 7.

$$I(u) = \begin{cases} C_j & \text{if } p(c = C_j|u) \geq T \\ OOS & \text{otherwise} \end{cases} \tag{7}$$

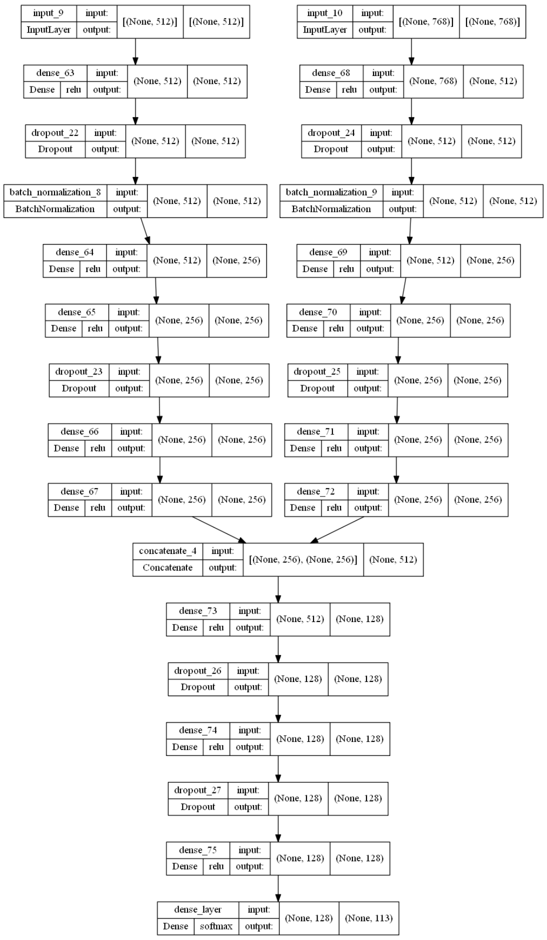

Figure 2: The model's architecture

It is noteworthy that the threshold $T \in [0, 1]$ is calibrated on the validation set without utilizing any OOS examples. Also, the original OOS examples from the training dataset were not employed during the training and validation phases of the DETER framework. Instead, unselected intents and synthetic outliers were utilized as OOS examples. Furthermore, the threshold was fine-tuned using the provided validation set.

4. Experimental Setup

We now present DETER's experimental findings for intent classification. The main objective is to assign the proper intent labels to utterances in the test set consisting of in-domain (known) and out-of-scope (unknown) intents. We separately report the approaches' macro *F1-score* performance for the known and unknown intents. We compare the performance of the proposed DETER framework with the state-of-the-art methodologies in the TEXTOIR framework (Zhang et al., 2021a).

4.1. Dataset

We evaluated the competing approaches on three real-world datasets, CLINC-150, Banking77, and Stackoverflow, as shown in the table 1.

CLINC-150 consists of 150 intent classes across ten domains designed for out-of-scope intent detection. The 1,200 OOS sample texts are split as 100 train, 100 validation, and 1,000 test (Larson et al., 2019). Banking77 is banking-specific data with 77 intents from 13,083 customer queries (Casanueva et al., 2020). Stackoverflow features 20 classes, each with 1,000 examples (Xu et al., 2015).

In this study, we consolidate all 1,200 OOS texts from CLINC-150 into the test set, as our framework does not utilize labelled OOS examples during training (Larson et al., 2019). Experiments conducted with 25%, 50%, and 75% of dataset intents were repeated five times each to ascertain model robustness and maintain consistent intents for each run.

4.2. Model Hyper-parameters

In our experimental setup, we set the batch size, number of epochs, and patience for stopping criteria on validation accuracy as 200, 1000, and 100, respectively. We maintain consistency with the epoch count and patience interval as the (K+1)-way method (Zhan et al., 2021) and replicate the experimental setup of the TEXTOIR framework (Zhang et al., 2021a). Specifically, we conduct ten experiments for each dataset, randomly selecting 25%, 50%, and 75% of intent classes from the training set as known classes for training, with the remainder reserved as unknown classes for testing. We utilize 1200 Out-Of-Scope (OOS) examples from the CLINC-150 dataset as OOS test samples for all datasets.

This range is chosen to analyze the impact of varying amounts of synthetic and open-domain outliers, with optimal results observed around 500 open-domain and 500 synthetic outliers. The (K+1)-way method employs the entire SQuaD 2.0 dataset, comprising nearly 100,000 data points, to generate open-domain outliers (Zhan et al., 2021), establishing the precedent for incorporating large numbers of outliers. While the SQuaD dataset is open-domain, the CLINC-150, Stackoverflow, and Banking77 datasets are specific closed-domain datasets, minimizing the likelihood of overlap. Additionally, the SQuaD dataset has been previously used as a proxy OOS example for intent classification (Zhan et al., 2021).

For the model optimization, we use $AdamW$ (Kingma and Ba, 2015) and *categorical cross entropy* as a loss function for the multi-class classification model.

We studied the impact of different outliers during model training using varied open-domain and syn-

Dataset	Intent	Training		Validation		Testing	
		Total	25%	Total	25%	Total	25%
Clinc-150	Known	15,000	3,800	3,000	760	4,500	1,140
	Unknown	0	11,200	0	221	1,200	3,360 + 1,200
Banking77	Known	9,003	2,119	1,000	234	3,080	760
	Unknown	0	6,884	0	221	1,200	2,320 + 1,200
Stackoverflow	Known	12,000	3,000	2,000	500	6,000	1,500
	Unknown	0	9,000	0	221	1,200	4,500 + 1,200

Table 1: Total number of examples and the computed examples of 25% intent ratios.

thetic outliers. Notably, on the validation set, we observed that excessive synthetic outliers adversely affect the open domain performance. Hence, the number of open-domain outliers varied from [50, 4000] and synthetic outliers from [50, 16000]. Further, for re-classification of the model predictions, the threshold T, based on the validation set, was set to 0.7.

5. Evaluation Results

We extensively compare our proposed DETER framework with existing OOS intent detection techniques. We consider 3 different case studies as:

Evaluating DETER Against State-of-the-Art

We compare our DETER framework with TEXTOIR state-of-the-art open intent detection techniques, namely `(i)` OpenMax (Bendale and Boult, 2016), `(ii)` DOC (Shu et al., 2017), `(iii)` ARPL (Chen et al., 2022), `(iv)` DeepUnk (Lin and Xu, 2019), `(v)` SEG (Yan et al., 2020), `(vi)` (K+1)-way (Zhan et al., 2021), `(vii)` ADB (Zhang et al., 2021b), `(viii)` MSP (Hendrycks and Gimpel, 2018), `(ix)` LOF (Breunig et al., 2000), `(x)` KNNCL (Zhou et al., 2022), `(xii)` MDF (Xu et al., 2021), and `(xiii)` DA-ADB (Zhang et al., 2023a).

For a fair comparison, each technique uses the same BERT model as backbone (Zhang et al., 2021a). Our study uses the TEXTOIR platform to standardize intent selection for training and testing (Zhang et al., 2021a). We used the same seeds as TEXTOIR for consistent results on benchmark datasets to ensure identical intents in DETER.

Results: The results from our proposed approach, evaluated at various known intent ratios (25%, 50%, and 75%) of the total CLINC-150 known intents, are presented in Table 2. Across all datasets and known intent ratios, the DETER framework consistently outperforms other methods in known and unknown intent detection. Notably, in the Clinc-150 and Banking-77 datasets, DETER's scores for unknown intents are particularly commendable. Some techniques like KNNCL and LOF face significant challenges with unknown intents in the Stackoverflow dataset. Notably, many current methods find it challenging to accurately classify known or unknown samples with different training intents. For instance, OSS detection is more efficient in 25% and 50% scenarios, while in-domain classification is superior for 75%. However, we outperform others in detecting in-domain and OOS in all scenarios.

DETER consistently achieves an impressive F1 score exceeding 90% for in-domain and out-of-scope (OOS) classification, even with varying training data, demonstrating its efficacy in data-scarce scenarios and adaptability to unfamiliar user inputs. The robust and superior performance of DETER underscores its capability. We observed that augmenting synthetic data volume enhances model performance, particularly under limited training data conditions (e.g., 25% scenario), attributed to improved delineation of OOS class boundaries with restricted in-domain datasets. Conversely, a balanced dataset proves more beneficial with extensive training data (e.g., 75% of intents).

Assessing DETER with Various Embeddings

Our investigation revealed that both the standalone USE model and unsupervised training of TSDAE using diverse datasets (Clinc-150, Askubuntu, Scidocs, and original Roberta) exhibited notably inferior performance compared to the proposed joint architecture of DETER.

Ablation experiments conducted on several dual encoder configurations (including USE only, TSDAE only without pre-training using Roberta, TSDAE with unsupervised learning on Clinc-150 combined with USE (DETER), TSDAE with unsupervised learning on Askubuntu combined with USE, and TSDAE with unsupervised learning on Scidocs combined with USE) using the Banking77 dataset are detailed in Table 3. From the results presented in Table 3, it is evident that the DETER framework (TSDAE with unsupervised learning on Clinc-150 combined with USE) consistently outperforms other configurations in both known and un-

Dataset	Model	Intent Ratio 25%		50%		75%	
		Known	Unknown	Known	Unknown	Known	Unknown
Clinc-150	(K+1)-way	74.02	90.27	81.52	84.25	86.72	79.59
	ADB	77.85	92.36	85.12	88.6	88.97	84.85
	ARPL	73.01	89.63	80.87	81.81	86.1	74.67
	DA-ADB	79.57	93.2	85.58	90.1	88.43	86
	DOC	75.46	90.78	83.84	87.45	87.91	83.87
	DeepUnk	76.95	91.61	83.3	87.48	86.57	82.67
	KNNCL	78.85	93.56	83.25	87.85	86.14	82.05
	LOF	77.77	91.96	83.81	87.57	87.24	82.81
	MDF	49.43	84.89	61.6	62.31	72.21	51.33
	MSP	51.02	59.26	72.82	63.71	83.65	63.86
	OpenMax	62.65	77.51	79.83	82.15	71.14	75.18
	SEG	46.67	59.22	62.57	61.34	42.72	40.74
	Our (DETER)	**92.19**	**98.42**	**92.02**	**96.26**	**92.15**	**92.51**
Banking-77	(K+1)-way	67.7	82.66	77.97	72.58	85.14	59.89
	ADB	70.92	85.05	81.39	79.43	86.44	67.34
	ARPL	62.99	83.39	77.93	71.79	85.58	61.26
	DA-ADB	73.05	86.57	82.54	81.93	85.93	69.37
	DOC	65.16	76.64	78.38	72.66	84.14	63.51
	DeepUnk	64.97	76.98	75.61	67.8	81.65	50.57
	KNNCL	65.54	79.34	75.16	67.21	81.76	51.42
	LOF	62.89	72.64	76.51	66.81	84.15	54.19
	MDF	44.8	85.7	64.27	57.72	75.47	33.43
	MSP	50.47	39.42	73.2	46.29	84.99	46.05
	OpenMax	53.42	48.52	75.16	55.03	85.5	53.02
	SEG	51.48	51.58	63.85	43.03	70.1	37.22
	Our (DETER)	**87.45**	**97.86**	**88.30**	**95.44**	**87.90**	**93.25**
Stackoverflow	(K+1)-way	50.54	52.23	70.53	51.69	81.2	45.22
	ADB	77.62	90.96	85.32	87.7	86.91	74.1
	ARPL	60.55	72.95	78.26	73.97	85.24	62.99
	DA-ADB	80.87	92.65	86.71	88.86	87.66	74.55
	DOC	56.3	62.5	77.37	71.18	85.64	65.32
	DeepUnk	47.39	36.87	67.67	35.8	80.51	34.38
	KNNCL	41.79	15.26	61.5	8.5	76.16	7.19
	LOF	40.92	7.14	61.71	5.18	76.31	5.22
	MDF	48.13	83.03	62.6	50.19	73.96	28.52
	MSP	42.66	11.66	66.28	26.94	81.42	37.86
	OpenMax	47.51	34.52	69.88	46.11	82.98	49.69
	SEG	40.44	4.19	60.14	4.72	74.24	6
	Our (DETER)	**88.16**	**97.35**	**88.82**	**94.71**	**88.16**	**90.35**

Table 2: Performance comparison (macro F1-score) of DETER framework for in-domain (known) and OOS (unknown) samples with varying amounts of known training intent ratios (25%, 50%, and 75%) for CLINC-150, Banking-77, and Stackoverflow datasets.

known intent detection tasks. Across all scenarios, DETER achieves an impressive F1 score exceeding 93% for unknown intents, representing up to a 10% improvement over the nearest embeddings. Similarly, for known intents, DETER achieves an F1 score surpassing 87%, exhibiting up to a 22% improvement over the nearest embeddings.

Model v/s Threshold Performance

Figure 3 illustrates a comparative analysis of the F1 performance of the model only (i.e., DETER without the re-classification threshold) and the complete DETER architecture on the CLINC-150 dataset for classifying known and unknown intents. Across all intent proportions, the model with threshold consistently outpaces the standalone version. This emphasizes the positive impact of the DETER framework on model performance. Notably, the model's aptitude in classifying unknown intents is commendable. At the 25% intent proportion, it reaches a striking 98.46% F1, demonstrating its proficiency in identifying out-of-scope utterances. Compared to using the model only, the DETER framework dis-

Embedding	Intent Ratio 25%		50%		75%	
	Known	Unknown	Known	Unknown	Known	Unknown
TSDAE (Clinc-150) & USE	**87.45**	**97.86**	**88.30**	**95.44**	**87.90**	**93.25**
TSDAE(Roberta) & USE	65.61	89.84	70.77	78.76	84.47	83.83
TSDAE (Askubuntu) & USE	62.43	88.52	76.23	84.61	81.77	82.09
TSDAE (Scidocs) & USE	56.32	88.39	74.77	84.53	82.17	81.45
USE_only	52.74	90.00	76.55	85.65	81.09	77.98

Table 3: Performance comparison (macro F1-score) of DETER framework for in-domain (known) and OOS (unknown) samples with varying amounts of known training intent ratios (25%, 50%, and 75%) for Banking-77 dataset using different embedding settings.

plays improved performance and presents lower standard deviations, indicative of model robustness. Although the performance gap narrows with increasing data, the model with the threshold consistently holds an edge. For known intents, the performance remains relatively stable across varying proportions, hovering around 91-92%, indicating model consistency in recognizing the intents it is trained on, irrespective of data volume. Similar behaviour was demonstrated across the other two datasets, Banking77 and Stackoverflow, underscoring the robustness and versatility of DETER.

Class Weights

To tackle the class imbalance issue (between in-domain and OOS training data), we evaluated DETER *with and without* class weights in this case study. We observed no significant performance difference in the two settings. For example, in the 25% proportion of CLINC-150, the best F1 results obtained with class weights are 82.07% for the known and 94.32% for the unknown. Furthermore, without class weights, the performances are 82.10% for known and 94.20% for unknown classes.

6. Discussion

Our research presents DETER, a framework for robust out-of-scope user utterance detection. DETER offers a comprehensive approach by combining dual text encoders (USE and TSDAE), a branched deep learning architecture, synthetic and open-domain outliers, and a threshold-based re-classification mechanism. We focus on formulating an efficient intent classification framework, leveraging the combined strength of USE and TSDAE sentence representations integrated with dense layers.

DETER's Performance

DETER's substantial improvements with limited labelled training sets (25% and 50% cases in Table 2) indicate its enormous potential in real-life applications. This is particularly significant in an ever-changing landscape of user interactions where adaptive systems are prominent.

Table 2 shows that the DETER framework performs best on three benchmark datasets (CLINC-150, Banking77, and Stackoverflow) and outperforms the leading models. In the CLINC-150 dataset, DETER outperforms the state-of-the-art models by 13%, 6%, and 3% for different known percentages (25%, 50%, and 75%) and 5%, 6%, and 6% for different unknown percentages (25%, 50%, and 75%). In the Banking77 dataset, DETER outperforms the state-of-the-art models by 16%, 7%, 2% for different known percentages and 11%, 13%, and 24% for different unknown percentages. Similarly, in the Stackoverflow dataset, DETER outperforms the state-of-the-art models by 7%, 2%, 0.5% for different known percentages and 6%, 5%, and 15% for different unknown percentages.

This performance improvement of DETER is attributed to its unique combination of Universal Sentence Encoder (USE) and Transformer-based Sequential Denoising Auto-Encoder (TSDAE), which offers nuanced sentence embeddings that adeptly bridge universal linguistic traits with domain-specific nuances. Its branched architecture bolsters feature extraction, clearly differentiating intent and OOS utterances. The framework also includes complementary components like synthetic outliers and threshold-based re-classification mechanisms, further enhancing its efficacy.

Synthetic and Open-Domain Outliers

Our findings show that synthetic outliers with higher values have a more noticeable impact than open-domain outliers, which extend beyond mere observation. This insight opens up new pathways for research to leverage synthetic data manipulation to enhance model robustness and performance.

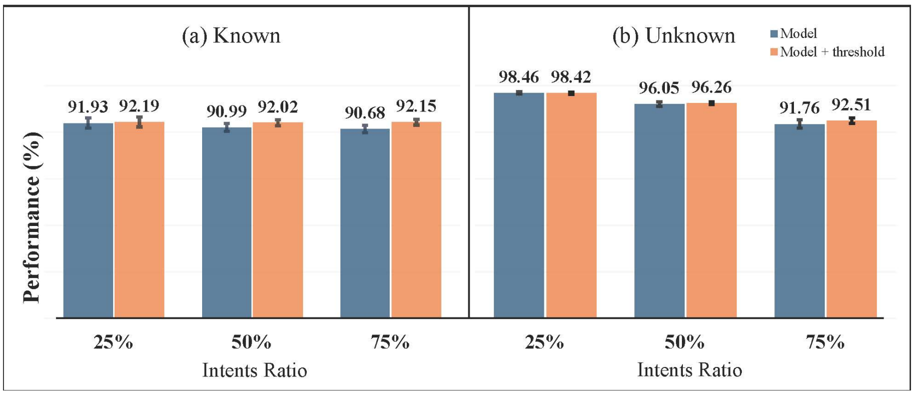

Figure 3: Performance comparison of the "Model only" versus the "Model with threshold (DETER)" on CLINC-150 dataset for both (a) known and (b) unknown intents across varying intent ratios . The error bars display the standard deviation across ten runs.

Efficiency, Scalability, and Applicability

One of the most salient aspects of this model is its remarkable computational efficiency. The DETER model boasts remarkable computational efficiency with only $1.5M$ trainable parameters, starkly contrasting the $125M$ in the model by Zhan et al. (2021). This streamlined design, which does not sacrifice performance, enhances scalability and flexibility for deployment across various platforms, even those with limited resources.

Performance and Robustness

Despite its simplicity of architecture, the model maintained robust performance, substantiating the efficacy of our approach. This performance parity with more complex models underlines our approach's potential to contribute significantly to ongoing research in deep learning and NLP.

Limitations and Future Work

Our assessment of the DETER framework encompassed 3 datasets, providing a broader evaluation spectrum. However, even with this multi-dataset approach, inherent data limitations persist. This emphasize the necessity of further exploration across even more diversified datasets, including multilingual datasets or languages other than English, to fully evaluate the framework's potential. Among future research avenues, few-shot learning is an intriguing domain poised to augment DETER's adaptability and efficiency.

7. Conclusions

This work introduces the Dual Encoder for Threshold-based Re-Classification (DETER), designed to detect out-of-scope user utterances in dialogue systems and by employing dual text encoders (USE and TSDAE), branched deep learning neural network architecture, synthetic and open-domain outliers, and a threshold-based re-classification mechanism, the DETER framework was rigorously evaluated using CLINC-150, Banking77, and Stackoverflow datasets. Our findings indicate a promising performance and superiority over the baseline. Through experimentation with synthetic outliers and open-domain data, we found that synthetic outliers exerted a more pronounced influence on the model's efficacy. This insight underscores the significance of adept outlier handling to augment the model's capability. However, introducing class weights had a negligible impact on the results.

While focusing on three benchmark datasets (CLINC-150, Banking77, and Stackoverflow) introduces certain constraints, the overarching conclusions underscore a promising direction for future research. The emphasis lies on optimizing efficiency, scalability, and applicability without sacrificing performance. DETER's evaluation of three benchmark datasets reveals its superiority, outperforming the state-of-the-art models.

This insight suggests a promising avenue for future research. Given that the TSDAE framework utilizes either Roberta or BERT models as its backbone, the possibility of training a multilingual TSDAE using multilingual data arises. This approach, mirroring the original TSDAE framework by Wang et al. (2021), could lead to the development of a multilingual DETER framework capable of handling diverse language datasets.

8. Acknowledgments

This project has received funding from Enterprise Ireland and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 847402.

9. Bibliographical References

Vibhav Agarwal, Sudeep Deepak Shivnikar, Sourav Ghosh, Himanshu Arora, and Yashwant Saini. 2021. Lidsnet: A lightweight on-device intent detection model using deep siamese network. In *20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1112–1117.

Abhijit Bendale and Terrance Boult. 2016. Towards open set deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, page 93–104, New York, NY, USA. Association for Computing Machinery.

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

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3952–3961. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. In *arXiv preprint arXiv:1803.11175*.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. 2022. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces. In *arXiv preprint arXiv:1805.10190*, pages 12–16.

Muhammad Fikri Hasani, Ford Lumban Gaol, Benfano Soewito, and Harco Leslie Hendric Spits Warnars. 2022. Deep learning and threshold probability for out of scope intent detection in task oriented chatbot. In *3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pages 322–327.

Dan Hendrycks and Kevin Gimpel. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *arXiv, 1610.02136*.

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

Dimitri Kartsaklis, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Mapping text to knowledge graph entities using multi-sense LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1959–1970. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316. Association for Computational Linguistics.

Shiyu Liang, Yixuan Li, and R. Srikant. 2020. Enhancing the reliability of out-of-distribution image detection in neural networks. In *arXiv preprint arXiv: 1706.02690*.

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

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA.

A. Emin Orhan. 2019. Robustness properties of facebook's resnext wsl models. In *arXiv preprint arXiv: 1907.07640*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688. Association for Computational Linguistics.

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

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

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online. Association for Computational Linguistics.

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

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021a. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 167–174, Atlanta, Georgia, USA.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021b. Deep open intent classification with adaptive decision boundary. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14374–14382.

Hanlei Zhang, Hua Xu, Shaojie Zhao, and Qianrui Zhou. 2023a. Learning discriminative representations and decision boundaries for open intent detection. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 31, pages 1611–1623.

Youwen Zhang, Xudong Wang, Linlin Wang, Ke Yan, and Huan Chen. 2023b. Out-of-scope intent detection with supervised deep metric learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. KNN-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141. Association for Computational Linguistics.