Label efficient semi-supervised conversational intent classification

Mandar Kulkarni, Kyung Kim, Nikesh Garera, Anusua Trivedi
Flipkart Data Science
(mandar.kulkarni, kyung.kim, nikesh.garera, anusua.trivedi)@flipkart.com

Abstract

To provide a convenient shopping experience and to answer user queries at scale, conversational platforms are essential for e-commerce. The user queries can be pre-purchase questions, such as product specifications and delivery time related, or post-purchase queries, such as exchange and return. A chatbot should be able to understand and answer a variety of such queries to help users with relevant information. One of the important modules in the chatbot is automated intent identification, i.e., understanding the user's intention from the query text. Due to non-English speaking users interacting with the chatbot, we often get a significant percentage of code mix queries and queries with grammatical errors, which makes the problem more challenging. This paper proposes a simple yet competent Semi-Supervised Learning (SSL) approach for label-efficient intent classification. We use a small labeled corpus and relatively larger unlabeled query data to train a transformer model. For training the model with labeled data, we explore supervised MixUp data augmentation. To train with unlabeled data, we explore label consistency with dropout noise. We experiment with different pre-trained transformer architectures, such as BERT and sentence-BERT. Experimental results demonstrate that the proposed approach significantly improves over the supervised baseline, even with a limited labeled set. A variant of the model is currently deployed in production.

1 Introduction

An automated conversational chatbot is essential to provide a seamless shopping experience and answer product-related questions at scale. An effective chatbot can assist and answer pre-purchase queries such as product specifications, offers, discounts, delivery time, and stock availability, as well as post-purchase queries such as exchange and return. Due to users from diverse backgrounds interacting with the chatbot and minimizing a human agent transfer, a chatbot should be able to understand and handle a variety of user queries.

One of the important ML components in the chatbot is automated intent identification, i.e., understanding the user's intention from the query text. Post the correct intent identification, an appropriate dialog-flow can be initiated. An incorrect intent prediction negatively affects the dialog-flow and, hence the overall user experience. Further, due to non-English speakers interacting with the chatbot, we observe a significant percentage of code-mix Hinglish queries (30%) and queries with grammatical errors, making intent detection even more challenging. Training a supervised intent classification model under such a scenario would require a large amount of manually tagged data. However, due to internet-scale operations, we have unlabeled query data available in a relatively large volume.

This paper proposes a simple yet competent Semi-Supervised Learning (SSL) approach for label-efficient intent classification. SSL has been proven effective in leveraging unlabeled data when only a small labeled set is available. Specifically, we train a transformer BERT model on a small labeled corpus along with a larger unlabeled query data. Starting with limited labeled queries, we explore supervised as well as unsupervised data augmentation techniques. For the supervised data augmentation, we explore MixUp (Zhang and Vaidya, 2021) and simple label preserving NLP augmentations (Ma, 2019). For training with unlabeled data, typically, SSL algorithms rely on an extra smoothness constraint which enforces the

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 96–102

model to make consistent predictions on an unlabeled sample and its slightly perturbed version. Moreover, it is observed that the type of noise/perturbation plays an important role and a trivial noise may not provide desired improvements (Xie et al., 2020). Recently, a simple noise such as dropout has shown promising results for contrastive learning (Gao et al., 2021). We explore label consistency loss with dropout noise to train the BERT model with unlabeled data. The model is trained with the linear combination of supervised and unsupervised loss components. One of the challenges with a limited labeled set is how to halt the training when the validation set is not available; otherwise, it may result in over-fitting. In our experiments, we perform the model updates till the *training loss* is converged. Interestingly, training with dropout label consistency loss is less prone to over-fitting even with no validation set. We also noticed that the choice of label consistency loss has a prominent effect on the accuracy. For warm starting the training, we experiment with pre-trained BERT and sentence-BERT architectures. Experimental results demonstrate that, over the supervised baseline, the intent classification accuracy can be boosted significantly with the proposed semi-supervised approach.

2 Related works

SSL approaches have been extensively studied in the literature. Instead of providing an extensive list of references, we only cite a few relevant prior works in this section. An extensive survey can be found in (Yang et al., 2021).

Unsupervised Data Augmentation (UDA) (Xie et al., 2020) has shown promising results for learning with unlabeled data along with a small labeled corpus. The idea is to enforce label consistency between two augmentations of the unlabeled sample. The authors also point out that the type of augmentation used significantly affects the accuracy of the model, and a trivial augmentation (such as adding Gaussian noise) may not lead to desired improvements. Recently, a contrastive learning approach that uses dropout noise has been shown to work well for self-supervised learning with textual data (Gao et al., 2021). Since dropout is inherently present in pre-trained transformer models, this provides a simple yet efficient method for data augmentation. Interpolation Consistency Training (ICT) (Verma et al., 2022) is a computationally efficient approach to train the model with SSL. ICT encourages the prediction at an interpolation of unlabeled points to be consistent with the interpolation of the predictions at those points. For classification problems, ICT moves the decision boundary to low density regions of the data distribution.

For the supervised classification, MixUp has been found to be an effective data augmentation technique (Jindal et al., 2020). MixUp is performed in the representation space for the text classification with transformers and is known to provide better regularization, and model calibration (Sun et al., 2020).

3 Proposed approach

In this section, we describe details of the dataset, loss functions experimented with, and model training.

3.1 Dataset

Our intent classification dataset consists of queries from the pre-defined set of 28 intents. The queries consist of pre-purchase as well as post-purchase user questions. For each intent, we have 250 manually labeled samples; hence, the train set comprises 7k labeled examples. As the test set, we use a manually tagged dataset of 7569 samples. Table 1 shows examples of the queries from the test set and corresponding ground truth intents. Note that the test set consists of code-mix Hinglish queries and queries with grammatical errors. For the unlabeled data, we use a query corpus of size ~925k obtained from the internal database. For all the queries (labeled and unlabeled), we convert them to lowercase and remove punctuation (if any). We do not apply any further pre-processing.

3.2 Loss functions experimented

We experiment with the following loss functions and their linear combination to train the model.

3.2.1 Supervised cross-entropy loss (l_s)

For a small set of labeled data, we use the standard supervised cross entropy loss for the

Samples	intent class
when will it be delivered if i order today this product satrday give me sir mujhe ye phone kab tak mile ga	delivery_time
when will we get discount it was 11000 near about 12000 at a time when it was offer phone ka price kab kem hoga	offers_and_discounts
is there debit card emi available emi process not full details show it option sorry sir card payment kaise karna hai	payment_options
is this boot washable sir this phone is good or but sir this phone prosser display kaise h ise mobile ki	product_spec
how much amount i will get into exchange of my mobile high what if the mobile i am replesing can be switched on mobile ka screen touch kharab hai exchange ho jaega	product_exchange
how to return my order my parking sensor not yet delvered humko black colour mila hai grey ke jagah	post_purchase

Table 1: Example queries and intent labels from the test dataset. Note that the test data contains code-mix Hinglish queries and queries with grammatical errors.

training. We use label smoothing while training where the smoothing parameter is set to 0.1. This loss function is included in all the experiments.

3.2.2 Supervised Grammar loss (l_{sg})

For the batch of labeled data, we add grammar augmentations to the input queries, such as spell errors and word swaps, to create additional train data (Ma, 2019). We use cross entropy loss and label smoothing for this.

3.2.3 Supervised MixUp loss (l_{sm})

The idea behind supervised MixUp is to create an additional labeled train set through linear interpolating of the features and corresponding one-hot labels. For the transformer models, MixUp is performed on the feature representations of the queries in the following manner.

$$\begin{aligned} \tilde{x} &= \lambda\ x_i + (1-\lambda)\ x_j \\ \tilde{y} &= \lambda\ y_i + (1-\lambda)\ y_j \end{aligned} \tag{1}$$

Here, $\lambda \sim U(0,1)$. x_i and x_j indicates the features from last hidden layer. We use cross entropy loss for this.

3.2.4 Unsupervised Dropout loss (l_{ud})

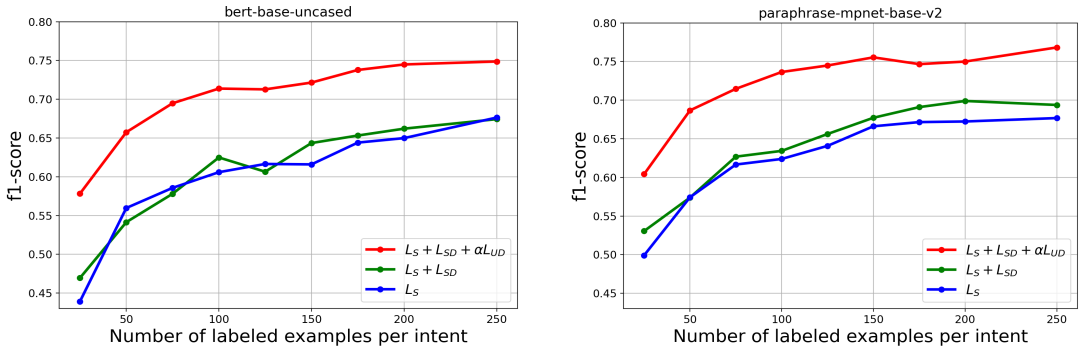
We use dropout noise for enforcing prediction label consistency to train the transformer model on unlabeled data. We sample a batch of queries from the unlabeled query corpus and make two independent forward passes through the transformer to obtain two label predictions. The label consistency loss is then calculated to minimize the distance measure D between these predictions.

$$l_{ud} = \mathbb{E}_{u \sim U(x)}\ D(p_\theta(y_1|u), p_\theta(y_2|u)) \tag{2}$$

Here, y_1 and y_2 indicate predicted labels for an unlabeled batch u. For D, we experimented with Cross Entropy (CE) and Mean-Square-Error (MSE) loss. For text classification, UDA uses round-trip back-translation as the data augmentation (Xie et al., 2020). They keep one copy of the network weights fixed while updating another copy. For the dropout, label predictions are calculated with the current network parameters, and the same is updated during training.

3.3 Training details

For the pre-trained BERT model, we use *bert-base-uncased* while for the pre-trained sentence-



(a) BERT results (b) Sentence-BERT results

Figure 1: F1-score comparison of BERT and Sentence-BERT results under different train settings.

BERT model, we use *paraphrase-mpnet-v2*. Both *bert-base-uncased* and *paraphrase-mpnet-v2* are 12 layers models with ~109M trainable parameters. For the BERT model, we use a feature corresponding to the *[CLS]* token from the last hidden layer (without tanh activation) as the query representation. For the sentence-BERT model, we use a mean-pooled representation of the token embeddings from the last hidden layer. The mean pooling uses an attention mask to avoid averaging representations from the padding tokens.

For the supervised losses (l_s, l_{sg}, l_{sm}), we use a batch size of 32, while for unsupervised loss (l_{ud}), we use a batch size of 96. We use AdamW optimizer with a constant learning rate of 1e-5. One major challenge with limited labeled sets is to halt the training without the validation set. In our experiments, we stop the training when the absolute difference in the train loss from the consecutive epochs remains below the threshold (ϵ) for a certain number of epochs (patience). In all our experiments, we use ϵ of 0.1 and patience of 5.

The models are trained under three different settings.

- Only with labeled loss, $L_S = l_s$
- With labeled loss (L_S) and supervised data augmentation loss, $L_{SD} = l_{sg} + l_{sm}$
- With labeled loss (L_S), supervised data augmentation loss (L_{SD}) and unsupervised dropout label consistency loss $L_{UD} = l_{ud}$. We use log probabilities along with MSE loss for L_{UD} and a weight factor α of 10 (to match the scales).

Figure 1 shows the comparison results for BERT and sentence-BERT models for varying number of labeled samples. We make a few observations from these results. Sentence-BERT works better than BERT, especially with a low number of labeled samples. Our findings align with the recent work demonstrating the effectiveness of Sentence-BERT for few shot learning (Tunstall et al., 2022). Supervised data augmentations (grammer + mixup) provide only a slight advantage over purely supervised baseline (Figure 1 (b)). We suspect it is happening due to over-fitting because of a small labeled corpus and lack of validation set to stop the training. We validate this hypothesis with an additional experiment, using some validation data to halt the training. Results are provided in the ablation study section 5.1. Unsupervised label consistency with dropout noise and MSE loss provides a significant advantage over the supervised baseline. Interestingly, even though the models are updated till the train loss is converged, training with this loss provides better regularization and is less prone to over-fitting. We also observe that the choice of unsupervised loss has a prominent effect on the accuracy. Section 5.3 in the ablation study shows the comparison results with different loss functions for l_{ud}.

Since Hinglish constitutes a significant percentage (30%) of queries, we specifically compared the performance of BERT and sentence-BERT models for Hinglish query classification. First, we detect Hinglish queries from the test set using an approach proposed in (Kulkarni et al., 2022) and calculate F1-score on these queries with the semi-supervised approach.

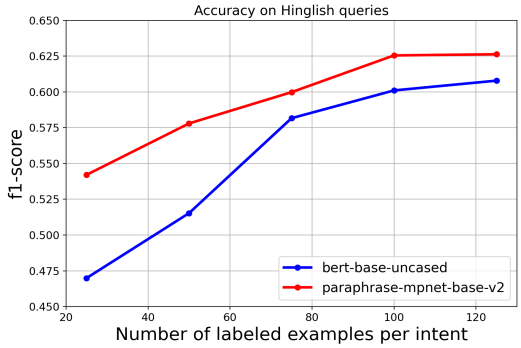


Figure 2: F1-score comparison of BERT and Sentence-BERT for Hinglish query classification.

Figure 2 demonstrates the result. We observe that sentence-BERT inherently provides better accuracies for Hinglish queries.

We also compare the Expected Calibration Error (ECE) on the test set for the BERT and sentence-BERT models. For this, we use the prediction result for the model trained on all the labeled samples. Table 2 shows the result. sentence-BERT achieves better calibration as compared to the BERT model.

setting	ECE
bert-base-uncased	0.0411
paraphrase-mpnet-v2	**0.0134**

Table 2: Comparison of Expected Calibration Error (ECE)

4 Comparison with Unsupervised MixUp approach

We compare the dropout label consistency approach with another SSL method: Unsupervised MixUp. Verma et al. (Verma et al., 2022) proposed a MixUp approach for training with unlabeled data. Feature MixUp is performed on the transformer representations for the two batches of unlabeled samples. For labels, MixUp on model predictions for the same unlabeled batches is used. We randomly sample two batches (u_1, u_2) from unlabeled queries and calculate their feature representation (x_1, x_2). The Unsupervised MixUp loss (l_{um}) is then calculated as follows.

$$l_{um} = \mathbb{E}_{u_1,u_2\sim U(x)}\, D(f_\theta(Mix_\lambda(x_1,x_2)), Mix_\lambda(f_{\theta'}(x_1), f_{\theta'}(x_2))) \quad (3)$$

As suggested in (Xie et al., 2020), for calculating the second term in the equation, we use a fixed copy (θ') of the network, and the update is applied to the current copy of the weights (θ). At the end of each epoch, a fixed copy is replaced with the current weights. The model is trained with supervised losses and the Unsupervised MixUp loss. We use MSE loss and α of 10. Figure 3 indicates the comparison result. Despite being simple, dropout label consistency performs better than Unsupervised MixUp. This could be because, at the start of the training, the predictions from the models may not be accurate. Hence, the updates to the model with Unsupervised MixUp loss are computed against noisy labels. On the contrary, the dropout consistency loss only enforces the smoothing constraint on the label predictions.

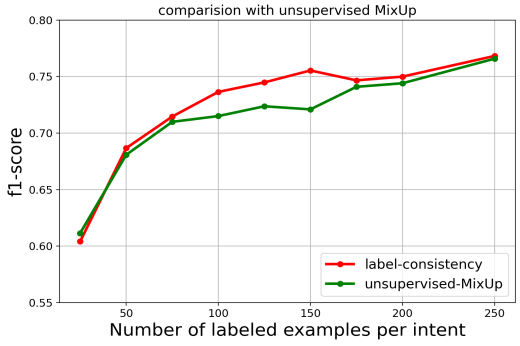


Figure 3: Comparison with Unsupervised MixUp.

5 Ablation study

In this section, we report ablation study results with different experimental settings.

5.1 Comparison of with and without validation loss monitoring

Since supervised MixUp provided only a slight improvement over the purely supervised baseline with sentence-BERT, we suspect that it is happening because of over-fitting since we do not have validation loss based stopping criteria during training. To confirm this, we conducted an additional experiment using a validation set (of size 8318) and halted the training when validation loss did not improve for five consecutive epochs. Figure 4 shows the F1-score comparison with and without validation monitoring. The plot indicates

that the supervised MixUp, when trained with a low number of labeled samples and without validation monitoring, is prone to over-fitting. Hence, it alone might not lead to good improvements for the limited labeled scenario.

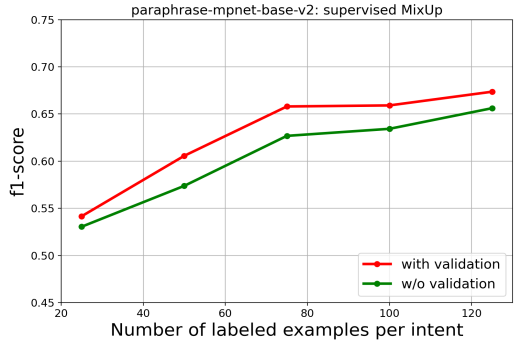


Figure 4: F1-score comparison for with and without validation loss monitoring. The result confirms that supervised MixUp is prone to over-fitting under low labeled data regime.

5.2 Choice of label consistency loss

We observed that the choice of loss used for dropout label consistency has a prominent effect on the model accuracy. Figure 5 shows the comparison of CE and MSE loss. For CE loss, we use α of 1, while for the MSE loss, α is set to 10 (to match the scales). It can be seen that the MSE loss consistently outperforms the CE loss.

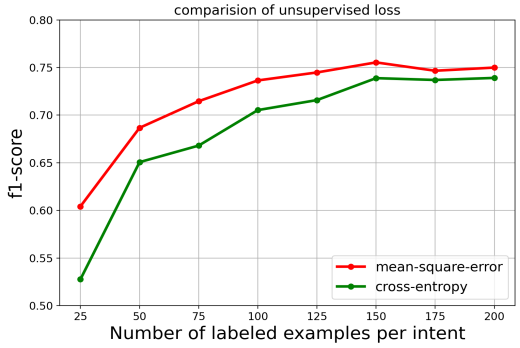


Figure 5: Effect of the choice of label consistency loss.

5.3 Effect of varying dropout probability

To understand whether model dropout probability affects the accuracy, we performed an experiment where we trained a sentence-BERT model with varied dropout probability. Sentence-BERT has a default dropout probability of 0.1. In this experiment, we set the dropout value to a lower (0.05) and a higher (0.2) value and trained the model with supervised and dropout label consistency losses. Figure 6 shows the resulting plot. We observe that increasing or decreasing the dropout probability does not significantly affect the model accuracy.

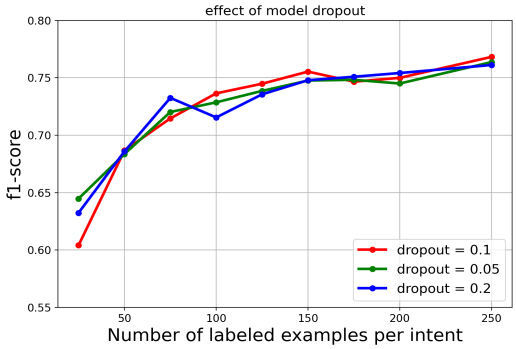


Figure 6: Effect of varying dropout probability.

6 Conclusion

This paper proposes a simple yet competent semi-supervised learning approach for label-efficient conversational intent classification. We trained different transformer models with labeled as well as unlabeled data. We explored supervised MixUp data augmentation for training with labeled samples, while for training with unlabeled samples, we experimented with label consistency loss with dropout. The results demonstrated that classification accuracy could be improved significantly over the supervised baseline with the proposed semi-supervised approach. Specifically, sentence-BERT was observed to perform better with a small number of labeled samples and even with code-mix Hinglish queries. Even without validation loss monitoring, it was noticed that training with dropout label consistency is less prone to over-fitting. Through the ablation study, we studied the effect of the choice of label consistency loss and dropout probability on the accuracy. Experimental results demonstrated the efficacy of the proposed approach. A variant of the model is currently deployed in production.

References

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings.

Amit Jindal, Dwaraknath Gnaneshwar, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Leveraging bert with mixup for sentence classification (student abstract). In *AAAI*.

Mandar Kulkarni, Soumya Chennabasavaraj, and Nikesh Garera. 2022. Study of encoder-decoder architectures for code-mix search query translation.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S. Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for nlp tasks.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. 2022. Interpolation consistency training for semi-supervised learning. *Neural Netw.*, 145(C):90–106.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2021. A survey on deep semi-supervised learning.

Wancong Zhang and Ieshan Vaidya. 2021. Mixup training leads to reduced overfitting and improved calibration for the transformer architecture.