Pre-trained Language Models with Limited Data for Intent Classification

Buddhika Kasthuriarachchy
School of Science, Engineering and Information Technology
Federation University Australia
Australia
b.kasthuriarachchy@federation.edu.au

Madhu Chetty
School of Science, Engineering and Information Technology
Federation University Australia
Australia
madhu.chetty@federation.edu.au

Gour Karmakar
School of Science, Engineering and Information Technology
Federation University Australia
Australia
gour.karmakar@federation.edu.au

Darren Walls
Global Hosts Pty Ltd
(trading as SportsHosts)
Australia
darren@sportshosts.com

***Abstract*—Intent analysis is capturing the attention of both the industry and academia due to its commercial and non-commercial significance. The rapid growth of unstructured data of micro-blogging platforms, such as Twitter and Facebook, are amongst the important sources for intent analysis. However, the social media data are often noisy and diverse, thus making the task very challenging. Further, the intent analysis frequently suffers from lack of sufficient data because the labeled datasets are often manually annotated. Recently, BERT (Bidirectional Encoder Representation from Transformers), a state-of-the-art language representation model, has attracted attention for accurate language modelling. In this paper, we investigate the application of BERT for its suitability for intent analysis. We study the fine-tuning of the BERT model through inductive transfer learning and investigate methods to overcome the challenges due to limited data availability by proposing a novel semantic data augmentation approach. This technique generates synthetic sentences while preserving the label-compatibility using the semantic meaning of the sentences, to improve the intent classification accuracy. Thus, based on the considerations for fine-tuning and data augmentation, a systematic and novel step-by-step methodology is presented for applying the linguistic model BERT for intent classification with limited data available. Our results show that the pre-trained language can be effectively used with noisy social media data to achieve state-of-the-art accuracy in intent analysis under low labeled-data regime. Moreover, our results also confirm that the proposed text augmentation technique is effective in eliminating noisy synthetic sentences, thereby achieving further performance improvements.**

***Index Terms*—intent classification, low data regime, language models, augmentation, semantic information, transfer learning**

I. Introduction

Nowadays, the community frequently expresses its wants and desires on social media platforms such as Twitter and Facebook. Understanding individual behavior through these contents, based on tasks such as sentiment analysis and opinion mining, has been an active area of research globally in the last decade [1]. The content-based intent analysis, aiming to identify the behavioral intention of users, falls within the domain of Natural Language Understanding (NLU). It has captured the attention of both the industry and academia due to its commercial and non-commercial significance, including linking buyers and sellers [2], identifying intentional behavior of seeking or offering help [3], and detecting malicious intents regarding sexual assaults [4]. The research being reported is part of a major project for a Melbourne start-up company Sportshosts[1] for classifying the intent of those international travelers to attend live team sports as a spectator while undertaking cultural tourism. For this, the tweets available in the public domain are being used as a data source to identify individuals who are interested in cultural tourism. Since the number of labeled tweets available is limited, any available method should be able to cope up with this limitation.

Currently, various classifier designs and techniques, incorporating the complexities of the automated intent classification task, have been reported using both heuristic methods and machine learning strategies. Recently, Hollerit et al. [2] proposed a binary classification method to identify the commercial intent of a tweet, applying supervised learning models using word n-grams and part-of-speech n-grams as features. However, the method fails to capture the semantic representations of the words. Pandey et al. [4] presented and evaluated an intent classification model for twitter posts using semantic features with the help of a convolutional neural network. However, the method uses only static word representations, and the model architecture makes it difficult to disregard the noise and focus on its relevance [5]. Most of these approaches [2], [4], [6] leverage bag-of-words representations, or static embeddings learned from shallow neural networks limiting these techniques since they suffer from the absence of dynamic representations of the words in a sentence. The dynamic representation is crucial as it enables understanding the human intentions.

In the recent past, word embedding [7] has become popular

[1]https://www.sportshosts.com/

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

among as a de facto starting point for representing the meaning of words. However, static methods such as Word2Vec [8], GloVe [9], and FastText [10] generally generate fixed word representations in a vocabulary, and hence these techniques cannot easily be adapted to a contextual meaning of a word. Recent discoveries of dynamic pre-trained representations such as ELMo, deep contextualized word representation [11], and BERT (Bidirectional Encoder Representations from Transformers), a language modeling framework [12] produce dynamic representations of a word based on the context. They capture many facets of language relevant for downstream tasks, such as long-term dependencies, hierarchical relations, and context to provide superior performance [13], [14]. Deep learning techniques with superior algorithms and complex architectures that leverage the contextual meaning of the words [7], [15], [16] can significantly improve the learning abilities. However, this performance depends, to a great extent, on the massive volume of labeled training data plays in making these deep learning models successful.

In our research that we are working related to cultural tourism, since we have meager resources to train the model, pre-trained language models such as BERT can prove to be suitable candidates for the effective transfer of natural language understanding into the intent analysis task. It may, however, be noted that the success of these dynamic representation models is heavily dependent on various factors. For intent analysis, the labeled datasets being often manually annotated may suffer significantly from a lack of accurately labeled training data imposing a major challenge in real-world scenarios. Further, different organizations may be interested in entirely different intent categories. Moreover, the same organization may look for a diverse set of intent categories over time. Additionally, organizations are keen to detect user intent from social media platforms due to its potential commercial value. However, social media data being noisy and diverse, it creates further challenges.

To address the problem of scarce labeled-data, Wang et al. [6] proposed a graph-based semi-supervised learning approach by using a tiny portion of labeled-data for model training. The nodes of the proposed graph are composed of words (intent-keywords) extracted from the tweets and assumes that tightly connected instances are likely belonging to the same class. However, intent keyword extraction limits the model from differentiating homographs efficiently. Transfer learning has also been employed to deal with the lack of sufficient labeled data for model training. For example, Dint et al. [17] proposed a framework of transfer learning based on CNN to classify implicit consumption intentions. They introduced a method to transfer the knowledge learned from a source domain to a target domain using a mid-level sentence representation learned using static representations generated from the Collobert and Weston (C&W) model [18]. Similarly, Pedrood and Purohit [3] introduced a novel approach of transfer learning using Sparse Coding feature representation to classify help intents into seeking, offering classes against the rest, during disasters. They efficiently transferred the knowledge for intent behavior previously learned from past disaster data. Compared to the other text classification research, only limited classifier designs and methods for content-based intent analysis have been reported in recent years.

However, to the best of our knowledge, no research is reported to apply the contextualized word representation to analyze the user intent and that too under limited availability of labeled data. The work reported in this paper is the first effort in this direction. In this paper, to overcome the challenge of capturing the contextual meaning of words and in particular intent-related information from small-scale noisy texts, we propose an inductive transfer learning with pre-trained language models for intent classification. Further, a novel *semantic data augmentation* is presented for augmenting the text data to boost the performance of intent classification models by preserving the label compatibility using the semantic meaning of the sentences. Experiments conducted with the published tweet dataset [6] reveal that the proposed method outperforms the current state-of-the-art graph-based semi-supervised approach to infer the intent categories in a low labeled data regime.

The rest of this paper is organized as follows. Section II provides relevant background information related to intent analysis, language modeling, and transfer learning. Section III introduces the proposed transfer learning approach and the mechanism for fine-tuning BERT, and also the novel semantic data augmentation technique. Section IV presents various experimental results on fine-tuning the BERT model with different settings for data augmentation, including the proposed approach. The main results are also discussed in this section. Finally, Section VI presents the conclusion.

II. Background

In this section, we elaborate on the relevant background information related to the proposed research. We first discuss the intent analysis and its unique characteristics. Next, we present BERT, the state-of-the-art deep bidirectional language representation model [12], which is an important component of the language modeling framework. This is followed by an explanation related to inductive transfer learning and its significance in the context of pre-trained language models.

A. Intent Analysis

The intent in the simplest term can be defined as a purpose for action. The intent analysis is the idea of identifying intentions present in textual content and recognizing a corresponding intent category for every action indicative of intent in a particular text [19]. Intent classification primarily attempts to capture a plausible future outcome [19] and is different from well-known text mining, such as opinion or sentiment classification, where they approximate the current state. For example, the sentence *"I like the color of iphone7"* reflects a positive sentiment, but no intention exists. In contrast, the sentence *"I want to buy an iphone7"* shows a firm buying intention in the near future. Therefore, verbs and keywords in a piece of text were considered essential features to identify

the intent. Hence term based intent analysis [2], [6], [20] has been a popular approach to detect intent.

B. Pre-trained BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is the first fine-tuning based language presentation model that achieves state-of-the-art performance on a broad suite of sentence-level and token-level tasks, outperforming many task-specific architectures [12]. BERT architecture includes a multi-layer bidirectional Transformer [5] and an attention mechanism that learns contextual relations between words (or sub-words) in a text. The Transformer consists of two separate mechanisms - an encoder that processes the input and a decoder that generates a prediction for the task. Since BERT is designed to generate a language model, only the encoder mechanism is used.

BERT trained bidirectionally on a large corpus of unlabeled text, including entire Wikipedia and Book Corpus, allows its models to understand the meaning of a language more correctly. Thus, it could be used for various target tasks such as sentiment classification, intent detection effectively. Two pre-trained BERT models were first introduced, i.e.,"$\text{BERT}_{\text{BASE}}$", that includes 12-layer bidirectional Transformer encoder block with 768 hidden units and 12 self-attention heads and also a "$\text{BERT}_{\text{LARGE}}$" consisting of 24-layer bidirectional Transformer encoder blocks with 1024 hidden units and 16 self-attention heads.

The processes of the tokenization of an input sentence for the BERT model involves splitting the input text into a list of tokens that are available in the vocabulary. To deal with the words not available in the vocabulary, BERT uses a technique called byte-pair-encoding (BPE) [21] based WordPiece tokenization [22]. The "$\text{BERT}_{\text{BASE}}$-uncased" version of the BERT models convert all the words of an input sentence to lowercase and uses a vocabulary of 30,522 words.

The input layer representation is a summation of WordPiece embeddings [22], positional embeddings, and the segment embedding. Since Transformers do not encode the sequential nature of an input sentence, positional embedding is used to introduce a temporal property. Segment embedding is used to distinguish a sentence pair, and it has no impact on a task based on a single sentence such as text classification. A special classification embedding ([CLS]) is prefixed as the first token of a sentence, and a special token ([SEP]) is appended as the final token. The final hidden state corresponding to the [CLS] token is used as the aggregate sequence representation for classification.

C. Inductive Transfer Learning

Transfer learning refers to the improvement of learning of a particular task by infusing the knowledge from prior learnings of a related task. Indeed, transfer learning has been playing an important role in many NLP applications [8], [23], and the learning strategy improves the performance on the target task leveraging the knowledge gained from a different but related concept or skill [24], [25]. Recently, Universal Language Model Fine-tuning (ULMFiT), introduced by Howard and Ruder [26], was seen as an effective inductive transfer learning method that can be applied to any task in NLP. However, the BERT model, with a similar approach, achieved superior state-of-the-art results [12].

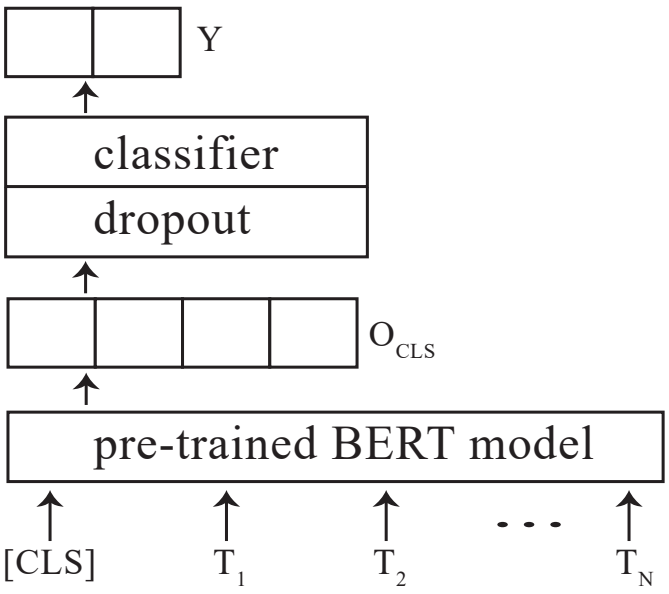

Fig. 1. The architecture of the BERT model extended for multi-class classification. T_i represents the WordPiece tokens of an input sentence. [CLS] is the special token introduced for classification tasks. O_{CLS} is the final hidden state corresponding to [CLS]. Y is the classification probability vector.

III. Proposed Approach

We propose to leverage the above-stated transfer learning to improve the performance of scarcely labeled intent classification tasks. To do this, a novel semantic data augmentation method generates synthetic sentences to tackle the small labeled data problem while preserving the label compatibility. Next, we focus on applying, for the first time, fine-tuning of a pre-trained BERT model for transfer learning and fine-tuning the model to classify intent using limited labeled data.

A. Semantic Data Augmentation

Due to the limited availability of labeled datasets, a gap exists between the amount of labeled and unlabeled data, thereby resulting in a tendency for the model to overfit the limited labeled data and underfit the unlabeled data.

To address this, we generate additional (i.e., synthetic) data via the transformation of a specific Tweet. These Available sentences are augmented without violating their meaning by applying the back-translation strategy (translating from English to any other language and then back to English) [28], which generates new semantically appropriate sentences that preserve the meaning of the original sentence - thereby synthesizing more data. While generating synthetic sentences, it is necessary to ensure that the synthetic sentence preserves the semantic similarity with the original sentence to reduce the risks of introducing a label noise.

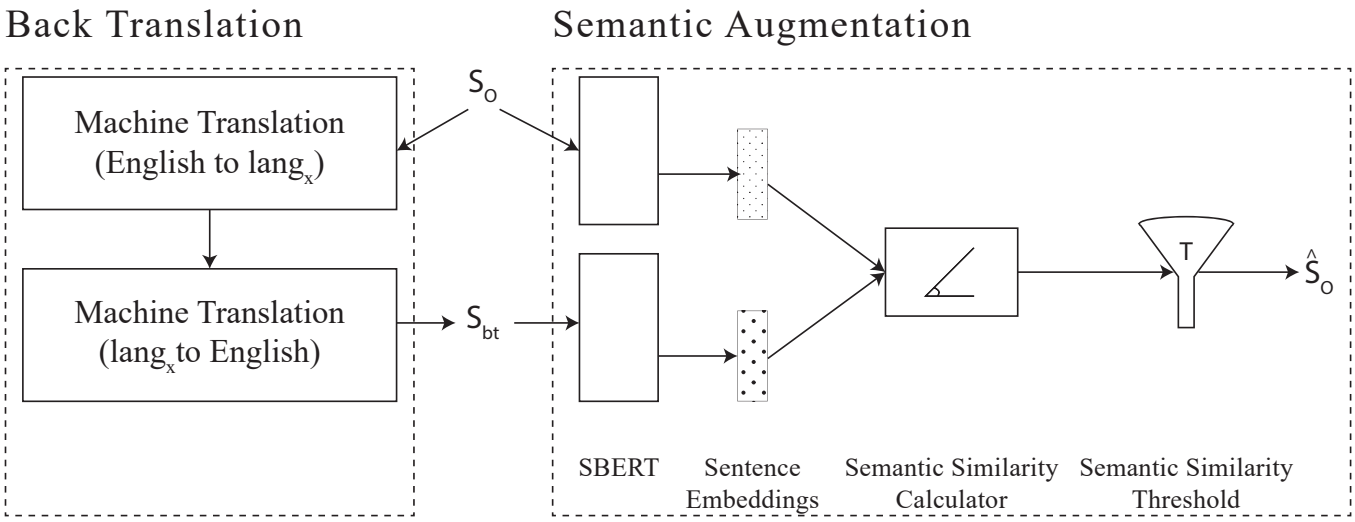

Fig. 2. Semantic data augmentation using back translation and sentence similarity filtering based on sentence embeddings extracted from the Sentence-BERT (SBERT) [27] model

A schematic diagram illustrating the proposed semantic data augmentation method is shown in Fig. 2. The key components of this semantic data augmentation architecture are back translation, sentence embedding, and similarity threshold. These are discussed next.

1) Back-translation: A new synthetic sentence is obtained by applying back-translation, which translates a tweet in English into any target language, $lang_x$, and then re-translates it back into English. The chosen target languages are those that belong to the Indo-European language family (i.e., same language family as English.) so that multiple target languages can be used effectively. For this research, the three target languages chosen are, namely, German, French, and Italian.

2) Sentence Embedding: While the diversity of words and sentences is essential, the semantic textual similarity of the sentences is also crucial to underpinning performance improvement, especially if the meaning of the sentence exists in the downstream segment. The proposed method leverages the semantic textual similarity (STS) to reduce the distortion or changes in the meaning of the original sentence. For this, we use the Sentence-BERT (SBERT), an existing built-in algorithm with the pre-trained BERT network and which uses the Siamese and triplet network structures to derive semantically meaningful sentence embedding [27].

3) Similarity Threshold: Let

S_o - source sentence

S_{bt} - synthetic sentence generated using a second language For data augmentation, we propose to use only those sentences which are semantically meaningful, by comparing the sentence embedding of the original sentence ($\vec{S_o}$) with the transformed example $\vec{S_{bt}}$. S_{bt} is considered a valid sentence ($\hat{S}_o$) if and only if $\cos(\vec{S_o}, \vec{S_{bt}}) \geq T$, where T is a chosen threshold value for semantic similarity.

To determine the similarity threshold level, we propose a novel method based on the probability density function of the cosine-similarity scores between S_{bt} obtained using all the target languages and corresponding original sentence S_o as depicted in Fig. 4. The method uses p^{th} percentile (π_p) to determine a set of candidate threshold values $\{T_{min}, T_p\}$, where $p \in 15, 25, 50$. T_p is calculated using Eqn. (1), whereas the minimum threshold value T_{min}, based on the standard interquartile range (IQR) rule (i.e., 1.5 x IQR rule), is calculated using Eqn. (2).

$$T_p = \int_{-\infty}^{\pi_p} f(x)\, dx \tag{1}$$

$$T_{min} = T_{25} - 1.5(T_{75} - T_{25}) \tag{2}$$

Since the augmentation dataset is composed of transformed sentences using multiple languages, a question then arises whether the different threshold values per target language might be more effective to identify semantically meaningful sentences. This approach is cumbersome and may be costly when we use a large number of target languages. However, the proposed approach is meaningful only if the behavior of the target languages are approximately similar to each other.

B. Fine-tuning

An intent classification, viewed as a multi-class classification problem with a predefined set of intent categories, can be accurately modeled using BERT. As shown in Fig 1, each tweet can be fed into the BERT model after tokenizing the tweet into WordPiece tokens $T = [[CLS], T_1, T_2, ..., T_N]$, to obtain the output $O = [O_{cls}, O_1, O_2, ..., O_N]$.

By leveraging the hidden state of its first special token ([CLS]), denoted $O_{cls} \in \mathbb{R}^H$, where H is the number of hidden units in the BERT model, the intent of each sentence S_i is predicted [12] as:

$$Y^i = \text{softmax}(WO^i_{cls} + b) \tag{3}$$

The only new parameters to be added [12] during the fine-tuning are for the classification layer $W \in \mathbb{R}^{K \times H}$ and also for $b \in \mathbb{R}^K$, where K is the number of classifier labels and H is the number of hidden units. Further, a dropout layer is added before the classification layer, with the dropout probability set to 0.1. This extension of the BERT model for multi-class classification is shown in Fig. 1.

To train the model, first, the standard Softmax function is applied to normalize the output of the classification layer $Y \in \mathbb{R}^K$ into a probability distribution of K probabilities. Then, the model is fine-tuned simultaneously, by considering all the parameters of BERT along with the classification layer weights W for minimizing the negative log-likelihood objective function.

C. Methodology

Based on the above considerations for fine-tuning and data augmentation, following systematic and novel step-by-step methodology is proposed for applying the linguistic model BERT for intent classification with limited data available.

1) Translate a source sentence (S_o) to a second language and then back to English. Multiple target languages can be used to generate multiple synthetic sentences from a given S_o. Let S_{bt}, be a synthetic sentence generated, and A_i be a set of S_{bt} generated by each second language i for a given set of source sentences.
2) Obtain deep contextualized word representation vector S_o^e for each source sentence and S_{bt}^e for corresponding back-translated sentences from A_i using Sentence-BERT (SBERT), a modification of the pre-trained BERT network that use Siamese and triplet network structures to derive semantically meaningful sentence.
3) Compute the cosine-similarity C $(-1 <= C <= 1)$ between the sentence embedding pair (S_o, S_{bt}) for all the sentences in A_i.

   The steps 4)-7) below determine the semantic similarity threshold.
4) Let us propose a hypothesis that the two probability density distributions being compared are equal. Verify the equality of the probability density functions of the semantic similarity scores of synthetic sentences in each A_i against each other and with reference to the probability density functions of the semantic similarity scores of all the back-translated sentences A' based on Eqn. 4 below.
$$A' = \bigcup_{i=1}^{n} A_i \tag{4}$$
5) The null and alternative hypothesis for the Kolmogorov–Smirnov Test can be formally stated as
$$H_0 : f(C) = f_0(C) \text{ for all } C$$
$$H_1 : f(C) \neq f_0(C) \text{ for at least one } C.$$
   Apply the two-sample Kolmogorov-Smirnov test (K-S test) [29], as it is sensitive to deviations in both location and shape of the empirical cumulative distribution functions of the two samples.
6) If enough evidence is unavailable to identify any difference between the probability density distributions of the mixture of all back-translated sentences A' and back-translated sentences generated by a target language A_i, then determine a global threshold value for all the target languages based on the probability density distribution of all back-translated sentences as depicted in Fig. 4, without applying individual threshold values for each second language.
7) Apply a threshold, T, on C to retain only the back-translated sentences semantically close to the source sentence. T is a hyper-parameter of the proposed semantic data augmentation model.
8) Finally, fine-tune the extended BERT model using the augmented dataset with the optimal values for task-specific properties of the BERT model (i.e., batch size, learning rate and the number of epochs) and the semantic similarity threshold T, identified during hyperparameter tuning.

IV. Experiments

Experiments are next carried out to study the proposed data augmentation technique to overcome the labeled-data scarcity and also to evaluate the effectiveness of pre-trained language models in intent classification. For this, the BERT model is initially fine-tuned with limited instances from each intent category for training. Next, the proposed technique for data augmentation is applied to this limited data set, and results for with and without data augmentation are compared.

TABLE I
Data Distribution

Intent category	Number of tweets
Career	159 (7.46%)
Event	321 (15.07%)
Food	245 (11.50%)
Goods	251 (11.78%)
Travel	187 (8.78%)
Trifle	436 (20.47%)
Non-intent	531 (24.92%)

A. Dataset Studied

As a benchmark, the dataset developed and studied earlier [6] is considered. This dataset contains 2130 manually annotated tweets across seven intent categories, as shown in Table I. Table I also shows the distribution of the dataset in these seven categories. Let

D_T - Entire labeled data comprising of 50 instances for each intent category were randomly sampled

D_V - Remaining labeled data left unused (to simulate limited data scenario)

We perform the hyper-parameter tuning for the BERT model using the five-fold cross-validation by taking only 10 random instances from D_T to train the model, and D_V is used for validation (similar to [6]).

B. Evaluation of BERT Fine-tuning

To fine-tune, it is recommended to set most of the BERT model parameters to the original values assigned during pre-training. However, as the optimal batch size, the learning rate, and the number of epochs, a range of possible values working well for specific text mining tasks are reported [12], [30]. To meet our requirements, we explore the optimal task-specific hyperparameters for an intent analysis under the limited availability of labeled data.

To derive the optimal batch size, learning rate, and the number of training epochs, we run an exhaustive search over the following task-specific hyperparameters of the extended BERT model. Apart from the range of possible values recommended for the hyperparameters [12], we also introduce values for tiny batch sizes (4 and 8 samples), since we are using a very small set of labeled-data for fine-tuning. The Adam optimizer ($\beta_1 = 0.9, \beta 2 = 0.999, L2$ weight decay of 0.01) [31] with learning rate warmup over the first 10% of the training steps, and linear decay of learning rate afterword (similar to [12]) is used to optimize the objective function. We use *accuracy* as the evaluation metric. The hyperparameter settings chosen for our experiments are:

- Batch size: 4, 8, 16
- Learning rate (Adam): 2e-5, 3e-5, 4e-5, 5e-5
- Number of epochs: 3, 4

To obtain the test set accuracies, the ten-fold cross-validation is carried out with 10 randomly sampled instances from D_T as training data, and using D_V as test data. The cross-validation prevents the model from overfitting the data. As the fine-tuning can sometimes be unstable due to the small training data set, several random restarts for each cross-validation experiment are performed.

For the experiments, we use the pre-trained BERT models provided by the PyTorch-Transformers library[2]. In our simulation experiments for very small training datasets, we observed the best performance to be consistently obtained for the mini-batch sizes 4 and 8. We also observed the optimization difficulties (a high variance in scores between the folds) associated with large batch sizes during the k-fold cross-validation due to overfitting. In contrast, small batch sizes achieved the best training stability, indicating improved generalization performance.

C. Ablation Study for Semantic Data Augmentation

For the data augmentation experiment using back-translation, we randomly chose three target languages: German, French, and Italian. We then apply Google translate API ("googletrans") to translate the randomly selected ten instances of each category from D_T. After removing the synthetic sentences that are exactly similar to the original sentence, the augmented dataset D_A is obtained. With this approach, for each training dataset sample, we generated three augmented datasets using the target languages chosen.

[2]https://github.com/huggingface/transformers

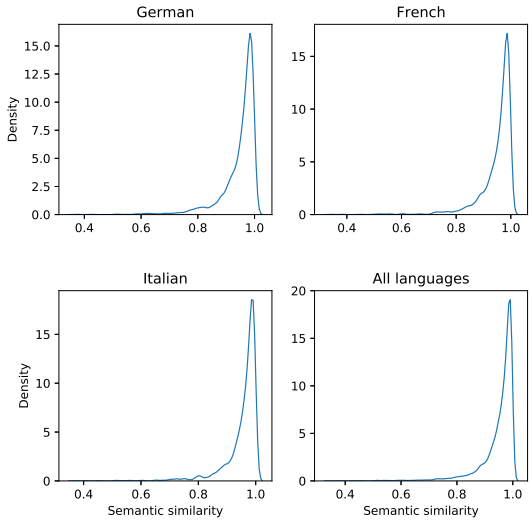

Fig. 3. Probability density distributions of semantic similarity scores of synthetic sentences

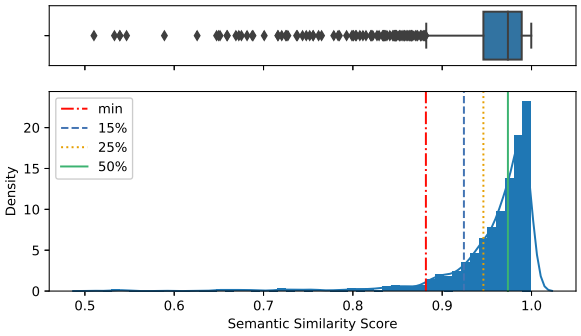

Fig. 4. Probability density distribution of semantic similarity of the mixture of all the back-translated sentences

Let each augmented dataset generated using different target languages be denoted as D_A^i, where $i \in 1, 2, 3$,

Sentence-BERT[3] is applied to generate sentence embeddings for original S_o and synthetic sentences S_{bt} in D_A^i. To evaluate the semantic similarity between S_o and S_{bt}, we have chosen the cosine similarity score (cosine of the angle between two embedding vectors), a widely implemented metric in information retrieval. The probability density distributions of the semantic similarity scores for back-translated sentences in each D_A^i and for the mixture of all the back-translated sentences are shown in Fig.3.

The hypothesis tests between the empirical distribution function of the mixture of all back-translated sentences and the distribution of the back-translated sentence generated using each of the target languages, revealed very high p-values for each KS test. This indicates weak evidence against the

[3]https://github.com/UKPLab/sentence-transformers

TABLE II

THE F1 RESULTS OF INDIVIDUAL CATEGORIES, AND MICRO-F1 AND MACRO-F1 OVER THE SEVEN CATEGORIES. THE MACRO-F1 WEIGHTS ALL THE CATEGORIES EQUALLY, WHEREAS THE MICRO-F1 WEIGHTS INDIVIDUAL TWEETS EQUALLY FAVOURING THE PERFORMANCE OF THE LARGE CLASS

Expt.#	Model	No. of back-tran	Threshold	Career	Event	Food	Goods	Non-intent	Travel	Trifle	Micro-F1	Macro-F1
1	Wang's	-	-	45.73	27.13	54.63	43.25	35.56	58.64	20.04	42.21	40.71
2	BERT (fine-tuning)	-	-	42.62	48.80	66.79	47.21	39.32	56.85	41.98	50.75	49.08
3	BERT+back-tr.	1	None	66.99	60.94	83.46	60.26	40.65	71.40	49.36	58.69	61.87
4	BERT+Sem. Aug.	1	T_{min}	65.76	61.86	84.20	61.59	48.74	75.20	45.59	60.14	63.28
5	BERT+Sem. Aug.	2	None	68.48	62.00	83.21	58.76	50.05	77.98	46.20	60.42	63.81
6	BERT+Sem. Aug.	2	T_{min}	68.30	64.45	84.54	60.48	53.88	74.57	45.22	**61.37**	**64.49**
7	BERT+Sem. Aug.	3	None	68.50	61.57	84.28	60.27	47.47	75.07	46.53	59.81	63.38
8	BERT+Sem. Aug.	3	T_{min}	69.03	63.33	84.17	59.50	50.28	74.38	46.82	60.65	63.93

null hypothesis, thereby failing to reject the null hypothesis. Therefore, a global threshold value is applied instead of all the back-translated sentences generated using different target languages, as shown in Fig. 4.

For fine-tuning, with the augmented dataset D_A^i, we follow the same model architecture as depicted in Fig. 1. We perform hyper-parameter tuning with five-fold cross-validation using D_A^i to obtain the optimal batch size, learning rate, number of training epochs and, additionally, the semantic similarity threshold of the proposed data augmentation technique. As depicted in Fig. 5, we obtained the best model performance on the training dataset for all D_A^i when the threshold value was set to T_{min}. For each D_A^i, the test accuracies were obtained using ten-fold cross-validation with ten randomly sample instances from D_T as training data, and D_V being used as test data.

To evaluate the effectiveness of our approach, we conduct several experiments. Table II presents the accuracies obtained by different strategies. Experiment 1 gives the baseline accuracies from Wang et al. [6]. Experiment 2 shows the performance of intent classification using the model based on BERT, as depicted in Fig.1, with the same dataset and a similar training set-up as [6]. Further, Experiments 3, 5, and 7 report the performance of our models, which were fine-tuned with D_A^i ($i = 1, 2, 3$ respectively), without applying any semantic similarity threshold T_p. However, for Experiments 4, 6, and 8, we apply the optimal semantic similarity threshold obtained during hyperparameter tuning to eliminate noisy synthetic sentences. As can be observed in Table II, we achieved 1.41%, 0.68%, and 0.55% average accuracy (Micro-F1) improvement for Experiment 4, 6, and 8, respectively, which signifies the effectiveness of the proposed semantic similarity threshold. This threshold, which controls the amount of noise removed from the synthetic dataset, contributes to the improvements. We observe that the highest accuracy with the proposed semantic data augmentation technique is with two back-translations with the similarity score threshold value of 0.8797 (T_{min}). Fig. 6 shows the performance of our approach with different threshold values for each D_A^i. As can be seen in Fig. 6, the model trained with semantically augmented training dataset outperforms the model trained with the full augmented dataset in terms of average test accuracy. Interestingly, we observe an overall drop in accuracy when there is an increase in the number of back-translations from two to three. This is possibly because the synthetic sentences are not providing any further diversity and variety to the training data, despite adding additional target languages, thereby resulting in the model overfitting the training data.

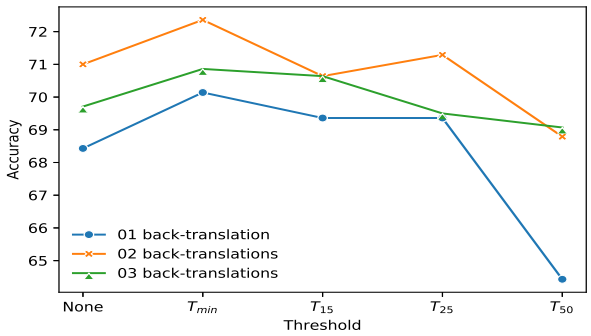

Fig. 5. Evaluation Accuracy

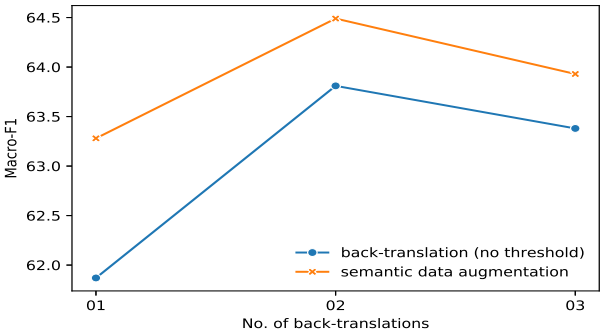

Fig. 6. Average Test Accuracy

V. DISCUSSION

The effectiveness of the proposed techniques becomes clear with the results of several experiments reported in the previous section. We note that the magnitude of the performance gains using the pre-trained language model is significant, even with minimal data being used for training. Despite using only ten

TABLE III
EXAMPLE SYNTHETIC SENTENCES AND SEMANTIC SIMILARITY SCORES

(a)	Original sentence	I need ice cream to put out fire -.-	
	Back-tr. (German)	I need ice fire extinguished -.-	0.72
	Back-tr. (French)	I need ice cream to extinguish the fire -.-	0.96
(b)	Original sentence	I need to hit up the mall .	
	Back-tr. (German)	I need the Mall beating.	0.83
	Back-tr. (Italian)	I have to hit the mall.	0.94
(c)	Original sentence	I want to buy i-phone .	
	Back-tr. (German)	I love shopping i-phone.	0.68
(d)	Original sentence.	I want chinese buffet for lunch .	
	Back-tr. (German)	I like Chinese buffet for lunch.	0.90
(e)	Original sentence	I should really get some sleeeep !	
	Back-tr. (French)	I should really sleep!	0.55
	Back-tr. (Italian)	I really should get some sleep '!	0.56
(f)	Original sentence	I would like to get a type writer ...	
	Back-tr. (German)	I want to get a typewriter ...	0.79
	Back-tr. (Italian)	I would like a typewriter ...	0.79
(g)	Original sentence	I should slerp with you . Hmm	
	Back-tr. (German)	I want to sleep with you. Hmm	0.37

instances from each intent category as training data, the BERT model, fine-tuned only with only four epochs, has performed remarkably well. It resulted in competitive accuracies against the more sophisticated semi-supervised learning models that require complex algorithms [6]. The fine-tuned BERT model obtains a significant absolute accuracy (Macro-F1) improvement of 8.4% over the state-of-the-art semi-supervised learning accuracy reported by Wang et al. [6]. Several experiments carried out clearly validate the effectiveness of the proposed technique. These results suggest that the pre-trained language models can have satisfactory performance even with noisy texts, and hence, they can be effectively utilized for other NLP applications having noisy data.

The examples (a)-(d) in Table III, back-translated sentences with corresponding semantic similarity scores proclaim that the proposed approach is more effective compared to the naïve back-translation. Interestingly, as observed in examples (a) and (b) in Table III, our approach was able to easily eliminate the meaningless back-translated sentences generated with German as the target language, while continuing to retain the meaningful and diverse synthetic sentences generated using French and Italian languages. Further, as observed in examples (c) and (d), the translations are acceptable in a general context. However, these synthesized sentences express an opinion rather than an intent [19]. The proposed semantic data augmentation technique maintains label compatibility in such situations.

In contrast, examples (e)-(g) in Table III, show valid synthetic sentences, but with low semantic similarity scores due to repeated sequential letters (e.g., sleeeep), incorrect word separations (e.g., type writer) and spelling mistakes (e.g., slerp) respectively. We may minimize the impact due to this by introducing pre-processing such as spelling correction and removing additional letters in a word with repeated sequential letters.

VI. CONCLUSION

With intent analysis, while the intended action can be inferred from the text, it may often require some contextual knowledge. The real-world applications of intent analysis are very much challenged by the scarcity of labeled data, hindering its successful application. This paper shows that significant improvement in the prediction accuracy of intent analysis can be achieved by transferring knowledge from the pre-trained language models to the intent analysis model. The pre-trained language model helps on two fronts: it allows the intent analysis model to understand the natural language efficiently and provides relevant knowledge learned from an extensive collection of unlabeled data that can be effectively used when the target task lacks enough labeled training data to identify important patterns. In this paper, we have shown that the use of BERT language modeling tool and a systematic step by step approach for the implementation of a novel semantic data augmentation technique can effectively infer the intent categories in a low labeled data regime. The proposed text augmentation reduces the noise introduced in synthetic sentences by maintaining the label-compatibility using the semantic similarity threshold enforced based on the cosine similarity score between original and the transformed sentences, thereby improving the overall accuracy. We are currently focussing on applying additional real-world datasets generated from relevant tweets for cultural tourism. The experiments with these datasets will help achieve further improvements with the proposed approach. It may be noted that the semantic similarity filtering technique is generic and can be applied in conjunction with any other text augmentation methods, especially when the semantic textual meaning of the specific task plays a key role in terms of accuracy.

The proposed model overcomes the limited data challenges from industry (including our project sponsor, SportsHosts) effectively and enables applying intent analysis successfully for such industry-related use cases.

ACKNOWLEDGMENT

This research is supported by Global Hosts Pty Ltd trading as SportsHosts, a Melbourne based company.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[2] B. Hollerit, M. Kröll, and M. Strohmaier, "Towards linking buyers and sellers: Detecting commercial intent on twitter," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13 Companion. New York, NY, USA: ACM, 2013, pp. 629–632.

[3] B. Pedrood and H. Purohit, "Mining help intent on twitter during disasters via transfer learning with sparse coding," in *Social, Cultural, and Behavioral Modeling*, R. Thomson, C. Dancy, A. Hyder, and H. Bisgin, Eds. Cham: Springer International Publishing, 2018, pp. 141–153.

[4] R. Pandey, H. Purohit, B. Stabile, and A. Grant, "Distributional semantics approach to detect intent in twitter conversations on sexual assaults," *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Dec 2018.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. USA: Curran Associates Inc., 2017, pp. 6000–6010.

[6] J. Wang, G. Cong, W. X. Zhao, and X. Li, "Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, ser. AAAI'15. AAAI Press, 2015, pp. 318–324.

[7] M. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 1499–1509.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *In EMNLP*, 2014.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[11] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[13] Y. Goldberg, "Assessing BERT's syntactic abilities," *CoRR*, vol. abs/1901.05287, 2019.

[14] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," *CoRR*, vol. abs/1906.04341, 2019.

[15] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 3261–3275.

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, nov 2018, pp. 353–355.

[17] X. Ding, T. Liu, J. Duan, and J.-Y. Nie, "Mining user consumption intention from social media using domain adaptive convolutional neural network," 2015.

[18] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," *CoRR*, vol. abs/1103.0398, 2011.

[19] M. Kröll and M. Strohmaier, "Analyzing human intentions in natural language text," in *Proceedings of the Fifth International Conference on Knowledge Capture*, ser. K-CAP '09. New York, NY, USA: ACM, 2009, pp. 197–198.

[20] A. Ashkan and C. L. Clarke, "Term-based commercial intent analysis," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 800–801.

[21] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, aug 2016, pp. 1715–1725.

[22] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[23] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[24] R. Vilalta, C. Giraud-Carrier, P. Brazdil, and C. Soares, *Inductive Transfer*. Boston, MA: Springer US, 2010, pp. 545–548.

[25] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 751–760.

[26] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339.

[27] N. Reimers and I. Gurevych, "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)*. Hong Kong, China: Association for Computational Linguistics, Nov 2019, pp. 3982–3992.

[28] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *CoRR*, vol. abs/1511.06709, 2015. [Online]. Available: http://arxiv.org/abs/1511.06709

[29] D. A. Darling, "The kolmogorov-smirnov, cramer-von mises tests," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 823–838, 1957. [Online]. Available: http://www.jstor.org/stable/2237048

[30] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham: Springer International Publishing, 2019, pp. 194–206.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." International Conference on Learning Representations (ICLR), 2015.