Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

LETS: A Label-Efficient Training Scheme for Aspect-Based Sentiment Analysis by Using a Pre-Trained Language Model

HEEREEN SHIM[1,2], DIETWIG LOWET[2], STIJN LUCA [3], AND BART VANRUMSTE[1].

[1]eMedia Research Lab & STADIUS, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[2]Philip Research, Eindhoven, the Netherlands
[3]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium

Corresponding author: Heereen Shim (e-mail: heereen.shim@kuleuven.be).
Joint last author: Stijn Luca (e-mail: stijn.luca@ugent.be) and Bart Vanrumste (e-mail: bart.vanrumste@kuleuven.be).

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 766139. This article reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains.

ABSTRACT

Recently proposed pre-trained language models can be easily fine-tuned to a wide range of downstream tasks. However, a large-scale labelled task-specific dataset is required for fine-tuning creating a bottleneck in the development process of machine learning applications. To foster a fast development by reducing manual labelling efforts, we propose a **L**abel-**E**fficient **T**raining **S**cheme (LETS). The proposed LETS consists of three elements: (i) task-specific pre-training to exploit unlabelled task-specific corpus data, (ii) label augmentation to maximise the utility of labelled data, and (iii) active learning to label data strategically. In this paper, we apply LETS to a novel aspect-based sentiment analysis (ABSA) use-case for analysing the reviews of the health-related program supporting people to improve their sleep quality. We validate the proposed LETS on a custom health-related program-reviews dataset and another ABSA benchmark dataset. Experimental results show that the LETS can reduce manual labelling efforts 2-3 times compared to labelling with random sampling on both datasets. The LETS also outperforms other state-of-the-art active learning methods. Furthermore, the experimental results show that LETS can contribute to better generalisability with both datasets compared to other methods thanks to the task-specific pre-training and the proposed label augmentation. We expect this work could contribute to the natural language processing (NLP) domain by addressing the issue of the high cost of manually labelling data. Also, our work could contribute to the healthcare domain by introducing a new potential application of NLP techniques.

INDEX TERMS

Active learning, Machine learning, Natural language processing, Neural networks, Sentiment analysis.

I. INTRODUCTION

Recently proposed pre-trained language models [1–3] have shown their ability to learn contextualised language representations and can be easily fine-tuned to a wide range of downstream tasks. Even though these language models can be trained without manually labelled data thanks to the self-supervised pre-training paradigm, large-scale labelled datasets are required for fine-tuning to downstream tasks. Data labelling can be labour-intensive and time-consuming creating a bottleneck in the development process of machine learning applications. Moreover, in real-world scenarios, the labelling scheme can be changed by adding or changing labels after deployment. Therefore, it is critical to be able to fine-tune the model with a limited number of labelled data to reduce manual labelling efforts and foster fast machine learning applications development.

One of the possible solutions is to apply active learning to reduce manual labelling efforts. Active learning is an algorithm designed to effectively minimise manual data labelling by querying the most informative samples for training [4]. Active learning has been extensively studied [4, 5] and applied to various applications, from image recognition [6, 7]

to natural language processing (NLP) tasks [8, 9]. Even though active learning guides how to strategically annotate unlabelled data, it does not utilise the unlabelled data or labelled data for fine-tuning. For example, unlabelled data points can be used for self-supervised learning or already labelled data points can be further utilised during supervised learning, such as by using data augmentation techniques.

To not only effectively reduce manual labelling efforts but also maximise the utility of data, we propose a novel **L**abel-**E**fficient **T**raining **S**cheme, LETS in short. The proposed LETS integrates three elements as illustrated in Fig. 1: (i) a task-specific pre-training to exploit unlabelled task-specific corpus data; (ii) label augmentation to maximise the utility of labelled data; and (iii) active learning to strategically prioritise unlabelled data points to be labelled. In this paper, we apply LETS to a novel aspect-based sentiment analysis (ABSA) use-case for analysing the reviews of a mobile-based health-related program. The introduced health-related program is designed to support people to improve their sleep quality by restricting sleep-related behaviour. We aim to provide a tailored program by analysing reviews of individual experience. To the best of our knowledge, this is the first attempt to implement an automated ABSA system for health-related program reviews. To illustrate the success of the novel use-case, we have collected a new dataset and experimentally show the effectiveness of the proposed LETS with the collected dataset and a benchmarks dataset.

The main contributions of this paper include the followings:

- A novel use-case of natural language processing and machine learning techniques for the healthcare domain is introduced (Sec. III);
- A novel label-efficient training scheme that integrates multiple components is proposed (Sec. IV);
- A label augmentation technique is proposed to maximise the utility of labelled data (Sec. IV-B2);
- A new query function is proposed to search different boundaries with two uncertainty scores for active learning with the imbalanced dataset (Sec. IV-B3);
- A new evaluation metric for an ABSA system is proposed to correctly evaluate the performance of a system in the end-to-end framework (Sec. V-C).

II. RELATED WORK

A. ASPECT-BASED SENTIMENT ANALYSIS

ABSA is a special type of sentiment analysis that aims to detect opinion toward fine-grained aspects. Since ABSA can capture insights about user experiences, ABSA has been widely studied in various industries, from consumer product sector [10, 11] to service sector [12–15]. ABSA entails two steps: aspect category detection and aspect sentiment classification [16]. During the first step, Aspect Category Detection (ACD), a system aims to detect a set of the pre-defined aspect categories that are described in the given text. For example, in the domain of restaurant review, the pre-defined set of aspects

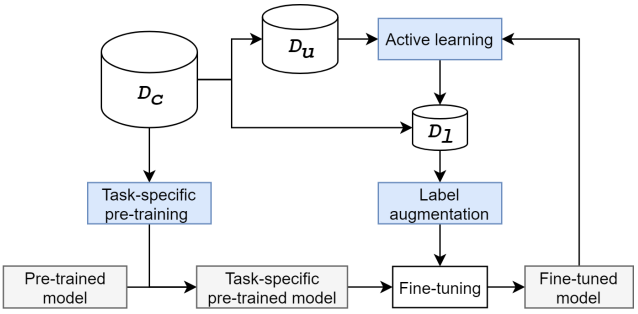


FIGURE 1. Overview of the proposed **L**abel-**E**fficient **T**raining **S**cheme (LETS). Task-specific pre-training utilises unlabelled task-specific corpus data set D_c. Label augmentation exploits labelled data set D_l. Active learning algorithm selects data from the unlabelled data set D_u for manual labelling.

can be {Food, Price, Service, Ambience, Anecdotes/Miscellaneous} and the task is to detect {Price, Food} out of the text "This is not a cheap place but the food is worth to pay". During the second step, Aspect Category Polarity (ACP), a system aims to classify a text into one of sentiment polarity labels (i.e., Positive, Negative, Neutral, etc) given a pair of text and aspect categories. For example, the task to produce a set of pairs, such as {(Price, Negative), (Food, Positive)} *given the set of ground truth categories* {Price, Food} and the text.

There has been significant improvement in ABSA systems over the past few years thanks to the recent progress of deep neural network (DNN) based NLP models, [10, 12, 13, 15, 17]. For example, Sun et al. [15] propose a Bidirectional Embedding Representations from Transformers (BERT) [1] based ABSA system by casting an ABSA task as a sentence-pair classification task. Even though this sentence-pair approach shows the state-of-the-art performance by exploiting the expanded labelled data set with sentence-aspect conversion[1] [15], it still requires a large amount of labelled data.

Later, Xu et al. [10] propose a post-training to utilise unlabelled corpus datasets to further train a pre-trained model for ABSA. The proposed post-training exploits both the general-purpose corpus dataset (i.e., texts from Wikipedia) and task-related corpus dataset (i.e., reading comprehension dataset) for the end task (i.e., review reading comprehension). Xu et al. [10] showed utilising multiple unlabelled corpus datasets can enhance the performance of the end task. Extensive studies on utilising unlabelled corpus for further pre-training showed that the importance of using domain-relevant data [18, 19]. However, domain-related corpus datasets for further pre-training are possibly not available in some domain (e.g., healthcare) because of privacy issue[2].

[1] As it is described in the original paper [15], a sentence s_i in the original data set can be expanded into multiple sentence-aspect pairs (s_i, a_1), (s_i, a_2), $\cdots$, (s_i, a_N) in the sentence pair classification task, with aspect categories a_n where $n \in \{1, 2, .., N\}$.

[2] For example, General Data Protection Regulation (GDPR) includes the purpose limitation principle mentioning that personal data be collected for specified, explicit, and legitimate purposes, and not be processed further in a manner incompatible with those purposes (Article 5(1)(b), GDPR).

B. ACTIVE LEARNING ALGORITHM

Active learning that aims to select the most informative data to be labelled has been extensive studied [4, 5, 20, 21]. The core of active learning is a query function that computes score for each data point to be labelled. Existing approaches include uncertainty-based [22, 23], ensemble-based [24, 25], and expected model change-based methods [4]. Thanks to their simplicity, uncertainty-based methods belong to the most popular ones. Uncertainty-based methods can use least confidence scores [8, 20, 26], max margin scores [27, 28], or max entropy scores [29] for querying.

One of the earliest studies of active learning with DNN is by Wang et al. [6] for image classification. They proposed a Cost-Effective Active Learning (CEAL) framework that uses two different scores for querying. One is an uncertainty score to select samples to be manually labelled. And the other is a certainty score to select samples to be labelled with pseudo-labels which are their predictions. Both scores are computed based on the output of DNN. Wang et al. [6] showed that the proposed CEAL works consistently well compared to the random sampling, while there is no significant difference in the choice of uncertainty measures, among the least confidence, max-margin, and max entropy.

However, other researchers claim that using the output of DNN to model uncertainty could be misleading [7, 30]. To model uncertainty in DNN, Gal and Ghahramani [30] proposed Monte Carlo (MC) dropout as Bayesian approximation that performs dropout [31] during inference phase. Later, Gal et al. [7] incorporated uncertainty obtained by MC dropout with Bayesian Active Learning by Disagreement (BALD) [32] to demonstrate a real-world application of active learning for image classification. Also, Shen et al. [8] applied BALD to an NLP task and experimentally showed that BALD slightly outperforms the traditional uncertainty method that uses the least confidence scores. The results from the large-scale empirical study by Siddhant and Lipton [9] also showed the effectiveness of BALD for various NLP tasks. Even though BALD outperforms the random sampling method, the differences between BALD and active learning methods with the traditional uncertainty scores (i.e., least confidence, max margin, and max entropy) are marginal [8, 9]. Also, BALD is computationally more expensive than the traditional methods because it requires multiple forward passes. Therefore, the traditional uncertainty scores are reasonable options when deploying active learning in a real-world setting.

Practical concerns on how to implement active learning in real-world settings include the issue that a model can perform poorly when the amount of labelled data is minimal [33]. This issue is referred to as the cold-start issue. Ideally, active learning could be most useful in low-resource settings. In practice, however, it is more likely that the model might work poorly with the limited number of labelled data at the beginning of active learning [34]. Therefore, introducing a component to ensures a certain level of performance with the limited labelled data is important to address the cold-start issue.

	Example
Free-text	I noticed that I was losing weight, but I missed the mid-afternoon caffeine boost most days. I slogged my way through work in the afternoon hours and missed the caffeine then, although I did sleep better.
Aspect	Energy: Negative Missing caffeine: Negative Sleep quality: Positive

TABLE 1. An example of aspect-based sentiment analysis based on the free-text user review of a health-related program.

III. ASPECT-BASED SENTIMENT ANALYSIS FOR HEALTH-RELATED PROGRAM REVIEWS

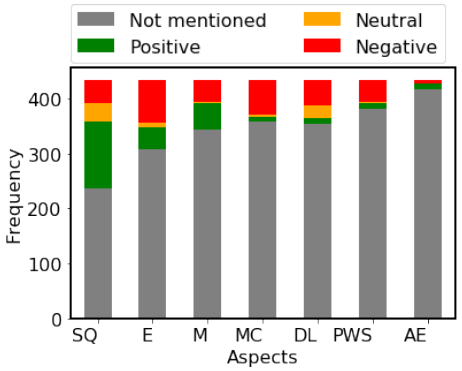
This section describes a mobile-based health-related program use-case that we call Caffeine Challenge. To conduct aspect-based sentiment analysis on the reviews of Caffeine Challenge, an experimental dataset is collected and annotated. The next subsections explain the details of the use-case, data collection protocol, and data labelling scheme with the initial data analysis result.

A. CAFFEINE CHALLENGE USE-CASE

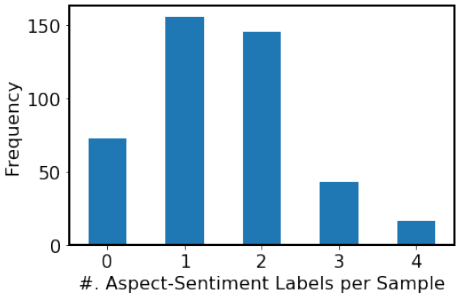
In this study, we introduce a health-related program that is designed to help people improve their sleep quality by restricting behaviour that might negatively affect their sleep quality. Having caffeinated beverage or desserts during the late afternoon and evening is selected as a target behaviour for this study. A challenge rule is restricting a caffeine intake after 13:00 for two weeks. During the program, participants use a mobile application to log their progress and receive notifications and recommendations of relevant information. At the end of the program, an in-app chatbot (conversational agent) asks about challenge experience and the participants are allowed to provide answers in free-text sentences. Our goal is to understand users' sentiments towards different aspects of the program by analysing the review data. To this end, we aim to develop an automated ABSA system for health-related program reviews as illustrated in Table. 1 where a system detects opinions (sentiment polarity) expressed towards multiple aspects. Since the ABSA system can capture detailed user opinions, it can be used to tailor the health-related program to individual users.

B. EXPERIMENTAL DATA COLLECTION

In the real-world machine learning application implementation process, multiple cycles on iterative development are often required: firstly, implementing a baseline model with experimental data and then gradually updating the model with real-world data. To develop the first version of the ABSA system, we conducted a pilot study with a semi-realistic dataset that is collected from an online survey via a crowd-sourcing platform (Amazon Mturk). At the beginning of the survey, an instruction containing details of the Caffeine Challenge (i.e., its purpose, goal, procedure, and consent form), is given to the survey participants. Then each participant has received a questionnaire regarding the experience of



(a) Sentiment class distribution per aspect category. Due to limited space, we use the following abbreviations: Sleep Quality (SQ), Energy (E), Mood (M), Missing Caffeine (MC), Difficulty Level (DL), Physical Withdrawal Symptoms (PWS), and App Experience (AE). Green, yellow, red, and grey bars indicate the number of samples with *Positive, Neutral, Negative*, and *Not mentioned* labels, respectively.



(b) Distribution of the number of aspect-sentiment labels per text excluding *Not mentioned* labels. The number of aspect-sentiment labels per sentence indicates the number of aspect categories mentioned in the sentence.

FIGURE 2. Annotation result of the collected Caffeine Challenge dataset. Sentiment class distribution per aspect category (a) and the number of aspect-sentiment labels per text (b) are shown.

the Caffeine Challenge. Then the participants have requested to answer the questions by imagining that they have done this challenge. In total, we recruited 1,000 participants and collected 12,000 answers and examples of collected data are shown in Appendix A.

C. DATA LABELLING

We annotated a random subset of the collected data for aspect-based sentiment analysis. Based on both health-related program and app development perspectives, seven different aspects are defined:

1) Sleep Quality (SQ)
2) Energy (E)
3) Mood (M)
4) Missing Caffeine (MC)
5) Difficulty Level (DL)
6) Physical Withdrawal Symptoms (PWS)
7) App Experience (AE)

Each aspect category is annotated with one of the sentiment values as follows: Positive, Neutral, Negative, and Not Mentioned. Not Mentioned class is introduced as a placeholder for an empty sentiment value. For example, when a sample does not describe any opinion towards a specific aspect, then it is labelled as Not Mentioned for that aspect category. A labelling scheme of each aspect category is given in Appendix B.

Fig. 2 illustrates annotation results and Fig. 3 shows the example of annotated data point. As it is shown in Fig. 2a, the majority of sentiment label within all aspect categories is an empty sentiment label (Not Mentioned). Some categories (Sleep Quality, Energy, and Mood) appeared more frequently compared to other categories (Missing Caffeine, Difficulty Level, Physical Withdrawal Symptoms, and App Experience). The former group is denoted as majority aspect

```
{
 'sentence':'I noticed that I was losing weight,
      but I missed the mid-afternoon caffeine
     boost most days. I slogged my way through
     work in the afternoon hours and missed the
      caffeine then, although I did sleep
     better.',
 'labels': {
     'sleep_quality': 'positive',
     'mood' : 'not_mentioned',
     'energy' : 'negative',
     'missing_caffeine': 'negative',
     'difficulty_level': 'not_mentioned',
     'physical_withdrawal_symptoms': '
         not_mentioned',
     'app_experience': 'not_mentioned',
 }
 }
```

FIGURE 3. An example of annotated data. Each annotated data point includes free-text and labels which are pairs of aspect category and sentiment class.

categories and the latter group is denoted as minority aspect categories. Fig. 2b shows the distribution of the number of aspect-sentiment labels per text, excluding Not Mentioned labels. It is observed that most of the annotated texts have either one or two aspect-sentiment labels and only a few have more than three aspect-sentiment labels.

IV. LABEL-EFFICIENT TRAINING SCHEME FOR ASPECT-BASED SENTIMENT ANALYSIS

We develop an automated ABSA system by utilising a pre-trained language model. Also, a label-efficient training scheme is proposed to reduce effectively manual labelling efforts. The following subsections will explain the ABSA system and the proposed label-efficient training scheme in detail.

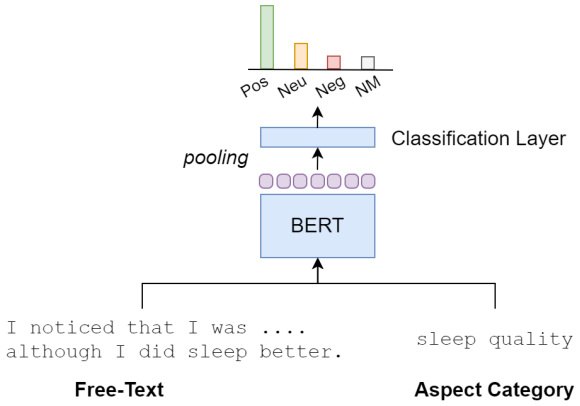


FIGURE 4. Illustration of aspect-based sentiment analysis (ABSA) as a sentence-pair classification by using Bidirectional Embedding Representations from Transformer (BERT).

A. ASPECT-BASED SENTIMENT ANALYSIS SYSTEM

Similar to the previous work by Sun et al. [15], we reformulate ABSA task as sentence-pair classification by using a pre-trained language model, BERT [1]. Fig. 4 illustrates a sentence-pair classification approach for ABSA. As shown in the figure, the proposed ABSA system produces the probability distribution over sentiment classes C, including polarised sentiment classes S (e.g., Positive, Neutral, Negative, etc) and an empty placeholder (e.g., Not Mentioned), for the given free-text sentence x_i and aspect category a_k. This formalisation allows a single model to perform aspect category detection and aspect sentiment classification at the same time. Also, adding an aspect category as the second part of input can be seen as providing a hint to the model where to attend for creating a contextualised embedding. Moreover, this formalisation allows expanding the training data set by augmenting labelled data, which will be explained in the following section (Sec. IV-B2).

Formally, an input is transformed into a format of $\boldsymbol{x}_i^k$ = [`[CLS]`, $\boldsymbol{x}_i$, `[SEP]`, $\boldsymbol{a}_k$, `[SEP]`], where $\boldsymbol{x}_i = [w_i^1, w_i^2, ..., w_i^{n_i}]$ is the tokenised i-th free-text, $\boldsymbol{a}_k = [a_k^1, a_k^2, ..., a_k^{m_k}]$ is the tokenised k-th aspect category in K aspect categories, and `[CLS]` and `[SEP]` are special tokens indicating classification and separation, respectively. Then the input is fed to the BERT model (f_θ) that produces contextualised embeddings for each token by using multi-head attention mechanism [1]. The contextualised embedding vector $\boldsymbol{e}_i^k \in \mathbb{R}^{d\times 1}$, corresponding to the classification token `[CLS]`, is used as the final representation of the given input $\boldsymbol{x}_i^k$. Then a classification layer projects $\boldsymbol{e}_i^k$ into the space of the target classes:

$$\boldsymbol{e}_i^k = f_\theta(\boldsymbol{x}_i^k) \tag{1}$$

$$\hat{\boldsymbol{y}}_i^k = \text{softmax}(\boldsymbol{W} \cdot \boldsymbol{e}_i^k + b) \tag{2}$$

where $\hat{\boldsymbol{y}}_i^k \in [0,1]^{|C|}$ is the estimated probability distribution over the sentiment classes C for the given free-text sample x_i and aspect category a_k pair, and f_θ, $\boldsymbol{W} \in \mathbb{R}^{|C|\times d}$, and $b \in \mathbb{R}^{|C|}$ are trainable parameters.

B. LABEL-EFFICIENT TRAINING SCHEME

One of the bottlenecks in developing an ABSA system with a pre-trained language model is to create a large-scale labelled task-specific dataset for fine-tuning which requires a labour-intensive manual labelling process. To mitigate this issue, we propose a **L**abel-**E**fficient **T**raining **S**cheme, which we refer as *LETS*. The proposed LETS consists of three elements to effectively reduce manual labelling efforts by utilising both unlabelled and labelled data. Fig. 1 illustrates the overview of the proposed LETS. The first element is task-specific pre-training to exploit the unlabelled task-specific corpus data. The second element is label augmentation to maximise the utility of the labelled data. The third element is active learning to efficiently prioritise the unlabelled data for manual labelling. The followings will describe the details of each element.

1) Task-specific pre-training

Task-specific pre-training is used to exploit the unlabelled task-specific corpus data. We adopt a novel pre-training strategy of Masked Language Modelling (MLM) from BERT [1] to train an Attention-based model to capture bidirectional representations within a sentence. More specifically, during the MLM training procedure, the input is formulated with a sequence of tokens that are randomly masked out with a special token `[MASK]` at a certain percentage p. Then the training objective is to predict those masked tokens. Since ground truth labels are original tokens, MLM training can proceed without manual labelling.

2) Label augmentation

Label augmentation is proposed to not only address the cold-start issue in active learning but also to maximise the utility of the labelled data. The proposed label augmentation algorithm multiplies the labelled data set by replacing aspect categories with similar words. This might look similar to common data augmentation techniques proposed by Wei and Zou [35] that includes synonym replacement, random insertion, random swap, and random deletion. Our method, however, modifies only the second part of the input (i.e., aspect category) while keeping the original free-text part. The proposed label augmentation technique is applied to pre-defined aspect categories with polarised sentiment classes S (e.g., Positive, Neutral, Negative, etc). Algorithm 1 summarises the proposed label augmentation technique.

3) Active learning

Active learning is used to prioritise the unlabelled data points to be manually labelled and added to the training pool. The core of active learning is a query function that scores the data points to use a labelling budget effectively in terms of performance improvement.

Algorithm 1: Label augmentation

Data: Labelled training set D_l, a dictionary of similar words per aspect category $Dict$, polarised sentiment classes S
Result: Augmented training set $\hat{D}_l$
$\hat{D}_l \leftarrow D_l$
for $d_l \in D_l$ **do**
  $txt \leftarrow$ getFreeText(d_l)
  $asps \leftarrow$ getAspects(d_l)
  for $asp \in asps$ **do**
    $senti \leftarrow$ getSentimentLabel(d_l, asp)
    if $senti \in S$ **then**
      $\hat{asps} \leftarrow Dict(asp)$
      for $\hat{asp} \in \hat{asps}$ **do**
        $\hat{d}_l \leftarrow (txt, \hat{asp}, senti)$
        $\hat{D}_l \leftarrow$ addData($\hat{d}_l$)
      end for
    end if
  end for
end for

Even though most of the existing active learning methods consider balanced datasets, one typical feature of a real-world dataset is that it can be imbalanced [36]. As it is shown in Sec. III-C, the collected dataset is also highly imbalanced: there are majority aspect categories that more often appear in the training set and minority aspect categories that less often appear in the training set. We observe that a fine-tuned ABSA model performs differently towards majority and minority aspect classes. For example, Fig. 5 illustrates the vector representations before the final classification layer[3] plotted into 2-dimensional space by using a dimensionality reduction algorithm [37]. From the figure, it is observed that the fine-tuned model can create distinctive representations between sentiment labels within the Sleep Quality aspect category, while the model fails to learn to differentiate data points among sentiment classes and empty sentiment class within the App Experience aspect category. This shows that a fine-tuned ABSA model performs relatively well towards majority aspect categories and its prediction is reliable, whereas a model works poorly towards minority aspect categories and it tends to fail to even detect the aspect categories.

Therefore, we propose two uncertainty measures for majority aspect categories and minority aspect categories, respectively:

$$u_{major} = 1 - Pr(\hat{y}_i^k = \arg\max_{c \in \mathcal{C}}(\hat{\boldsymbol{y}}_i^k) | \boldsymbol{x}_i^k) \quad (3)$$

$$u_{minor} = 1 - |Pr(\hat{y}_i^k = nm | \boldsymbol{x}_i^k) - \sum_S (Pr(\hat{y}_i^k = s | \boldsymbol{x}_i^k))| \quad (4)$$

$$= 1 - |1 - 2Pr(\hat{y}_i^k = nm | \boldsymbol{x}_i^k)| \quad (5)$$

[3]The fine-tuned model at the initial step of active learning experiment (Sec. V-D1) is used.

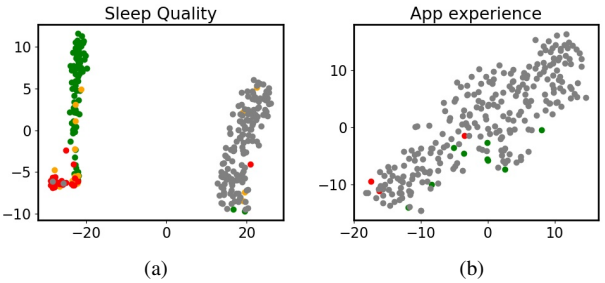


FIGURE 5. The final vector representations of inputs plotted in 2-dimensional space for Sleep Quality (a) and App Experience (b) aspect categories. Green, yellow, red, and grey colour indicate inputs with *Positive, Neutral, Negative*, and *Not Mentioned* sentiment labels, respectively. All data points were not used during the training phase.

where $Pr(\hat{y}_i^k = \arg\max_{c \in \mathcal{C}}(\hat{\boldsymbol{y}}_i^k) | \boldsymbol{x}_i^k)$ is the highest probability in the estimated probability distribution over sentiment classes given $\boldsymbol{x}_i^k$, nm refers *Not Mentioned*, and S refers a polarised sentiment classes set (e.g., Positive, Neutral, Negative, etc). u_{major} is the traditional least confidence score and u_{minor} is the margin of confidence score between an empty placeholder (i.e., Not Mentioned) and sum of other sentiment classes. As it is shown in (5), u_{minor} treats the model's prediction as binary classification result (i.e., *Not Mentioned* or *Mentioned*) producing high uncertainty scores when $Pr(\hat{y}_i^k = nm | \boldsymbol{x}_i^k)$ is close to 0.5. The intuition of introducing u_{minor} is allowing a model to focus on detecting whether the aspect category is mentioned or not. The proposed two uncertainty measures allow the model to search different boundaries during active learning: the boundaries where the model is uncertain about its aspect category sentiment classification result towards majority classes is described by u_{major}. And the boundary where the model is uncertain about aspect category detection result towards minority classes is described by u_{minor}.

Algorithm 2 shows the proposed LETS that integrates three elements. Firstly, a pre-trained model is further pre-trained with an unlabelled task-specific corpus data set. Then the task-specific pre-trained model is used for initialisation during active learning iterations. Active learning is repeated t times and each time a model is fine-tuned with the labelled data set that is augmented by the proposed label augmentation technique. At the end of each iteration step, n samples are queried from the unlabelled set for manual labelling. For querying, each Query function Q_{major} and Q_{minor} select $n/2$ samples where u_{major} and u_{minor} are the highest, respectively.

V. EXPERIMENTS

A. DATASETS

We evaluate the proposed method on two datasets. One is the custom dataset that we collected for the Caffeine Challenge use-case. The other is SemEval-2014 [16] task 4 dataset[4] that

[4]https://alt.qcri.org/semeval2014/task4/

Algorithm 2: Label-efficient training scheme (LETS)

Data: Pre-trained model M_{pt}, unlabelled task-specific corpus data set D_c, initial training set D_l, unlabelled training set D_u, total iteration t, labelling budget n, query function for majority categories Q_{major}, query function for minority categories Q_{minor}

Result: Fine-tuned model M_t, Labelled data set D_t

$M_{tspt} \leftarrow$ task-specificPre-train(M_{pt}, D_c,)

$i = 0$

$D_i \leftarrow D_l$

while $i < t$ & $|D_u| > 0$ **do**

  $D'_i \leftarrow$ augmentLabel(D_i)

  $M_i \leftarrow$ fineTune(M_{tspt}, D'_i)

  $d_{major} \leftarrow Q_{major}(D_u, M_i, n/2)$

  $d_{minor} \leftarrow Q_{minor}(D_u, M_i, n/2)$

  $D_{i+1} \leftarrow D_i$

  $D_{i+1} \leftarrow$ addData(addLabels($d_{major} \cup d_{minor}$))

  $D_u \leftarrow D_u - \{d_{major} \cup d_{minor}\}$

  $i+ = 1$

end while

is the most widely used benchmark dataset for aspect-based sentiment analysis.

1) Custom dataset: Caffeine Challenge

The custom dataset, which is described in Sec. III, is named as a Caffeine Challenge dataset. We annotate a random subset of the Caffeine Challenge dataset with 7 different aspect categories (i.e., Sleep Quality, Energy, Mood, Missing Caffeine, Difficulty Level, Physical Withdrawal Symptoms, App Experience) and 3 sentiment labels S ={Positive, Neutral, Negative} and an empty placeholder (i.e., Not Mentioned). The aspect categories distribution of the Caffeine Challenge dataset is imbalanced as described in Sec. III. Aspect categories are divided into subgroups of majority and minority aspect categories based on the frequency in a training set: {Sleep Quality, Energy, Mood} as majority aspect categories and {Missing Caffeine, Difficulty Level, Physical Withdrawal Symptoms, and App Experience} as minority aspect categories.

The unlabelled corpus data set are used for task-specific pre-training and the annotated data set is used for fine-tuning. Table 2 summarises the sizes of the Caffeine Challenge dataset used for the experiments. For task-specific pre-training, sentences from the unlabelled corpus data set are used. For the fine-tuning, 5-fold cross-validation splits are created at the sentence level and sentence-aspect pairs are used for training.

2) Benchmark dataset: SemEval

The SemEval-2014 task 4 dataset contains restaurant reviews annotated with 5 aspect categories (Food, Price, Service, Ambience, Anecdotes/Miscellaneous) and 4 sentiment

Data set	Sentence	S-A pairs
Unlabelled corpus	22,577	-
Training	325	2,275
Test	87	609
Total Fine-tuning	412	2,884

TABLE 2. Size of Caffeine Challenge dataset used for the experiments. Sentences from the unlabelled corpus data set used as the task-specific corpus data for task-specific pre-training. S-A pairs indicate sentence-aspect pairs and sentence-aspect pairs from the training set are used for fine-tuning.

Data set	Sentences	S-A pairs
Training	2,435	12,175
Test	609	3,045
Total	3,044	15,220

TABLE 3. Size of SemEval dataset used for the experiments. Sentences from the training set are used as the task-specific corpus data for task-specific pre-training. S-A pairs indicate sentence-aspect pairs and sentence-aspect pairs from the training set are used for fine-tuning.

labels S ={Positive, Neutral, Negative, Conflict[5]}. Since the SemEval dataset is also imbalanced, as illustrated in Appendix. C, we define majority and minority categories: {Food, Anecdotes/Miscellaneous} and {Service, Ambience, Price} as majority and minority aspect categories, respectively.

We used the original SemEval train set for the experiments to create 5-fold cross-validation splits. Table 3 summarises the size of SemEval dataset used for the experiments. For task-specific pre-training, sentences from the training set are used. For the fine-tuning, sentence-aspect pairs are created with an empty placeholder (Not Mentioned) for the sentences that do not contain a sentiment label towards specific aspect categories.

B. EXPERIMENTAL SETTINGS

1) Task-specific pre-training and fine-tuning

We use the pre-trained uncased BERT-base model as the pre-trained model (PT). The task-specific pre-trained model (TSPT) is created by further training the pre-trained model on the task-specific corpus data with the masked-language modelling (MLM) objective with masking probability $p = 0.15$. The TSPT is used to initialise the proposed method and the PT is used to initialise other methods during the active learning process. For fine-tuning, the final classification layer is added and entire model parameters are updated. More detailed implementation and hyperparameter settings are given in Appendix. D.

2) Label augmentation

Label augmentation multiplies the amount of labelled data by generating synthesised pairs of sentence and aspect categories by replacing aspect categories with similar words. The pre-defined dictionary containing a list of similar words

[5]The conflict label applies when both positive and negative sentiment is expressed about an aspect category [16]

Error type	Target	Prediction	Comparison
TP	TARG $\in S$	PRED $\in S$	TARG = PRED
NA	Not Mentioned	Not Mentioned	TARG = PRED
FN1	TARG $\in S$	Not Mentioned	TARG $\neq$ PRED
FN2	TARG $\in S$	PRED $\in S$	TARG $\neq$ PRED
FP	Not Mentioned	PRED $\in S$	TARG $\neq$ PRED

TABLE 4. Types of error used to compute aspect category sentiment classification (ACSC) scores. TP, NA, FN1, FN2, FP refer to true positive, not applicable, false negative type 1, false negative type 2, false positive, respectively. TARG and PRED refer to a target sentiment class and a predicted sentiment class where S is a set of polarised sentiment classes (e.g., Positive, Neutral, Negative, etc).

is used for label augmentation and label augmentation is applied to the only minority aspect categories to avoid inefficient augmentation. The pre-defined dictionaries are given in Appendix E.

3) Active learning

Active learning experiments are repeated 5 times with 5-fold cross-validation splits. At each fold, the initial labelled data set (i.e., seed data) is randomly selected from the training set at the sentence level and transformed into sentence-aspect pairs. For the Caffeine Challenge dataset, 20% of the training set (n=455) is used as seed data (D_l) and the remaining data is used as unlabelled data (D_u). For the SemEval dataset, 10% of the training set (n=1,220) is used as seed data (D_l) and the remaining data is used as unlabelled data (D_u). Active learning is iterated with 10 steps with a fixed labelling budget (n=|D_u|/10). At the initial iteration step (t=0), a model is trained on the seed data. During active learning steps, more data are iteratively added to the training set by selecting unlabelled data.

For comparison, we implemented BALD by using MC dropout [30], Cost-Effective Active Learning (CEAL) [6], least confidence scores, and random sampling. For BALD, we use the same approximation by Siddhant and Lipton [9] to compute uncertainty score as the fraction of models which disagreed with the most popular choice. The number of stochastic forward passes for BALD is set to 10. For CEAL, the least confidence score is used for calculating uncertainty and the threshold for pseudo-labelling is set to 0.05 with a decay rate of 0.0033. Since pseudo-labels are not included in the labelling budget, the active learning with CEAL can be terminated early when there is no more data for manual labelling. More details of these methods can be found in the original papers [6, 9].

C. EVALUATION METRICS

In this paper, we used two different metrics to evaluate the performance of an ABSA system. One metric is aspect category detection (ACD) and the other metric is aspect category sentiment classification (ACSC). Aspect category detection (ACD) is proposed by Pontiki et al. [16] and limited to evaluating aspect category detection ignoring the performance of aspect category sentiment classification. Aspect category polarity (ACP) metric is proposed to assess the sentiment classification performance of a system [16]. However, as it is mentioned in the previous study by Brun and Nikoulina [14], the ACP metric presumes the ground truth aspect categories and cannot be used to correctly evaluate an ABSA system end-to-end. To address this issue, we introduce a new metric of aspect category sentiment classification (ACSC) which is the modified version of ACP taking into account false aspect category detection results.

1) Aspect category detection (ACD)

ACD is used to evaluate how a system accurately detects a set of aspect categories mentioned in the input text. F_1 score is used which is defined as:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where precision (P) and recall (R) are:

$$P = \frac{|E \cap G|}{|E|}, R = \frac{|E \cap G|}{|G|}$$

where $|*|$ denotes the cardinality of a set *, E is the set of aspect categories that a system estimates for each input, and G is the set of the target aspect categories. Micro-F_1 scores are calculated at sentence-level and averaged over all inputs and macro-F_1 scores are calculated and averaged at aspect category-level.

2) Aspect category sentiment classification (ACSC)

ACSC is used to evaluate the performance of an ABSA system end-to-end. Since the proposed ABSA system produces multiple sentence-pair predictions for a single text input, the predictions are aggregated to compute (aspect, polarity) pairs at sentence-level while eliminating the pairs that contain Not Mentioned as a target as well as a predicted sentiment class. F_1 scores are calculated on the (aspect, polarity) pairs at aspect-level following:

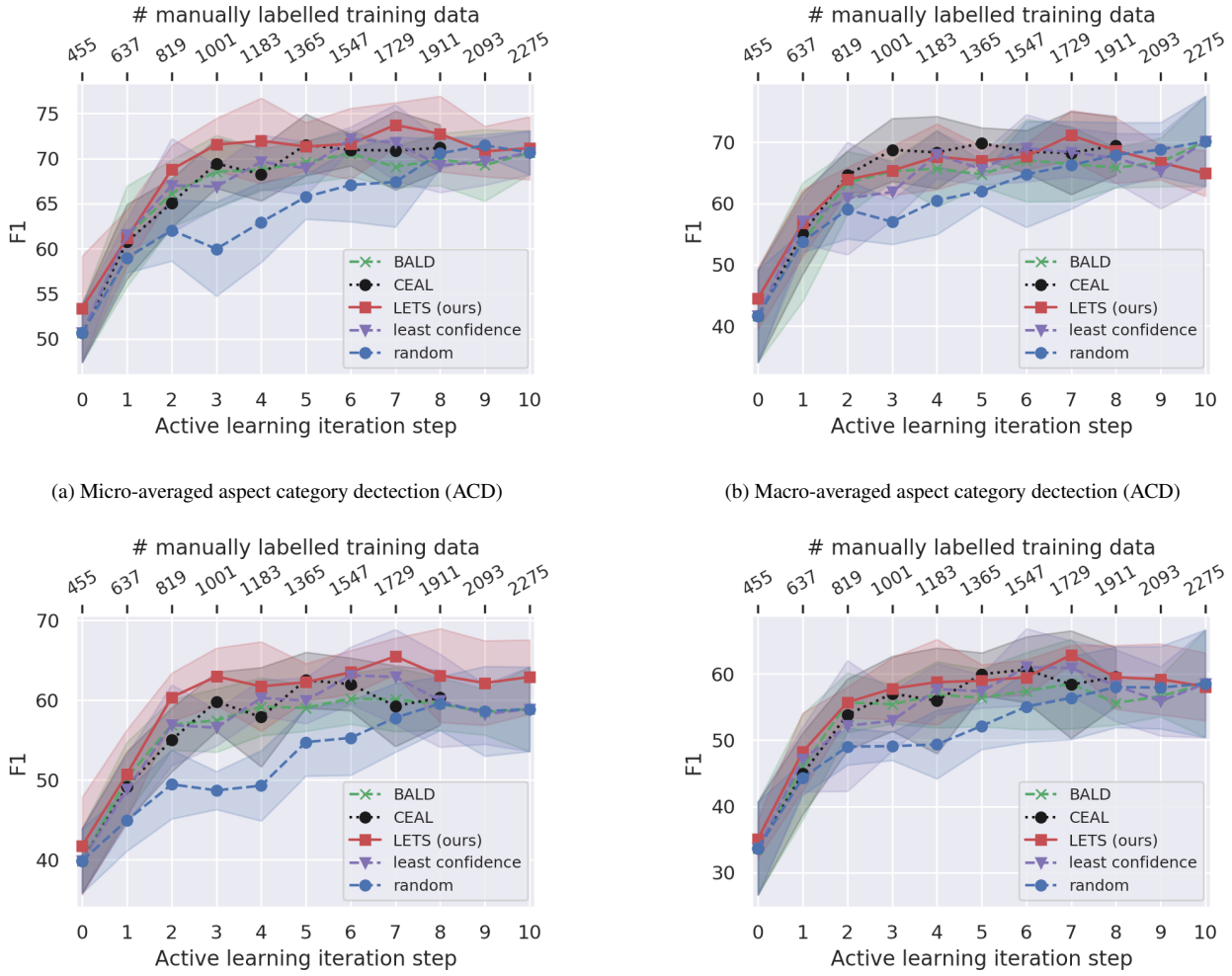
$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN1 + FN2}$$

where TP, FP, FN1, and FN2 are defined as in Table 4. Similar to ACD, both micro- and macro-averaged F_1 are used.

D. RESULTS AND ANALYSIS

1) Exp 1: Caffeine Challenge

Fig. 6 illustrates the active learning results with the Caffeine Challenge dataset. Active learning results in ACD metrics are illustrated in Fig. 6a and Fig. 6b. All active learning methods show better performance improvement than random sampling. It is observed that all models achieve much lower performances in macro-averaged scores than micro-averaged scores. These results show that the models perform worse towards minority aspect categories in the Caffeine Challenge dataset. In micro-averaged ACD score, LETS outperforms other active learning methods in general. In macro-averaged ACD score, CEAL achieves slightly better performance than



(a) Micro-averaged aspect category dectection (ACD)

(b) Macro-averaged aspect category dectection (ACD)

(c) Micro-averaged aspect category sentiment classification (ACSC)

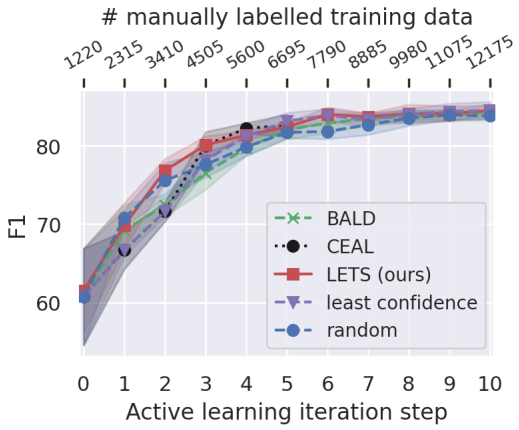
(d) Macro-averaged aspect category sentiment classification (ACSC)

FIGURE 6. Active learning results with the Caffeine Challenge dataset. Each line indicates averaged 5-fold results with standard deviation as shade. The bottom X-axis indicates the active learning iteration step and the top x-axis indicates the number of manually labelled training data. Y-axis indicates the performance score.

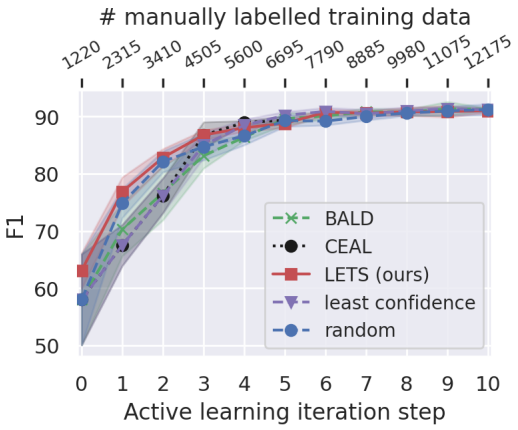
LETS. However, the ACD metrics are incomplete because they ignore sentiment classification results.

ACSC metric is proposed to address the limitation of the ACD metric and correctly evaluate the ABSA system end-to-end. Fig. 6c and Fig. 6d illustrate active learning results with the respect to the ACSC metrics. From the figures, it is observed that the performances of all models decrease compared to the observations from the ACD metrics. Similar to the results with the ACD metrics, LETS shows better performance improvement compared to other active learning methods. Specifically, from iteration step 0 to 1, the performance of LETS increases from 35.1% to 48.2%, while other method increase from 33.7% up to 47.1% in macro-averaged ACSC metric. The most significant difference is observed between LETS and random sampling. For example, random sampling achieves a similar performance of 48.2% at iteration step 2-4. Moreover, the difference between LETS and random sampling increases over iteration steps. The random sampling method at iteration step 6-7 and LETS at iteration 2 show similar performances in terms of macro-average ACSC metric. These results suggest that LETS can reduce manual labelling efforts 2-3 times better compared to the random sampling method. Also, LETS slightly outperforms other active learning methods at the beginning of the iteration step with the respect to the ACSC metrics. This result shows that the task-specific and the proposed label augmentation can contribute to better generalisability with the Caffeine Challenge data set.

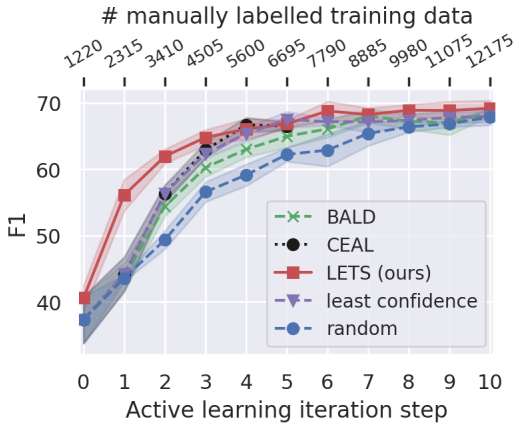
Performance differences between LETS and random sampling method are statistically significant (Wilcoxon signed-rank test with $p < .05$) from iteration step 1 to 7 and iteration step 2 to 5 in micro-and macro-averaged ACSC metrics, respectively. However, performance differences between LETS and active learning methods are not statistically significant ($p > .05$) throughout the entire iteration steps. In general, all methods show high variances of performances.



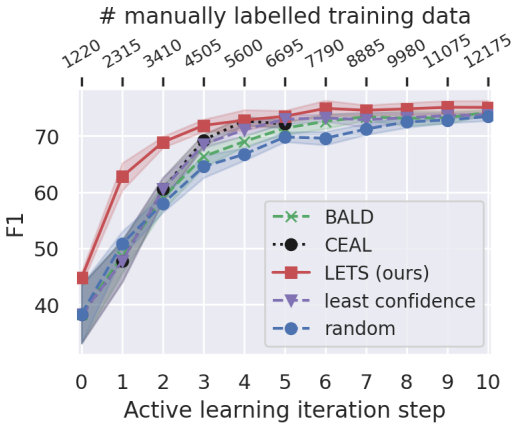
(a) Micro-averaged aspect category dectection (ACD)



(b) Macro-averaged aspect category dectection (ACD)



(c) Micro-averaged aspect category sentiment classification (ACSC)



(d) Macro-averaged aspect category sentiment classification (ACSC)

FIGURE 7. Active learning results with the SemEval dataset. Each line indicates averaged 5-fold results with standard deviation as shade. The bottom X-axis indicates the active learning iteration step and the top x-axis indicates the number of manually labelled training data. Y-axis indicates the performance score.

One interesting observation is CEAL achieves lower performances than LETS in terms of micro-averaged ACSC metric, especially in the later iteration steps. This is different from the observation from the micro-averaged ACD metric. A possible explanation for this is as follows: CEAL uses pseudo-labels. These pseudo-labels might not correct in terms of sentiment classes and errors might propagate throughout the iteration steps. Since the ACD metrics ignore sentiment classification results, this error might not be detected. Results with the macro-averaged ACSC metric show similar trends to the results with the macro-averaged ACD metric. These results suggest LETS slightly outperforms CEAL in terms of end-to-end evaluation metric.

2) Exp 2: SemEval

Fig. 7 illustrates the active learning results with SemEval benchmark dataset. Compared to the results with the Caffeine Challenge dataset, it is observed that the results with the SemEval dataset show less fluctuated learning curves in general. It is potentially because the SemEval dataset contains fewer aspect categories with more training data.

As illustrated in Fig. 7a and Fig. 7b, LETS shows slightly faster learning curves compared to other methods in terms of the ACD metrics. The random sampling method shows better learning curves compared to other active learning methods (i.e., BALD, CEAL, least confidence) in the ACD metrics. However, this does not imply that the random sampling method outperforms other active learning methods because the ACD metrics ignore sentiment classification results.

Fig. 7c and Fig. 7d show the active learning results in terms of the ACSC metrics. It is observed that the performances of all models decrease compared to the observations from the ACD metrics because the ACSC metrics consider sentiment classification results. From the figures, we can also see that the random sampling method achieves slower learning curves compared to the active learning methods. These results are opposite from the results with the ACD metrics and imply that the model trained with randomly sampled data tends to more misclassify sentiment labels.

In the ACSC metrics, it is observed that LETS substantially outperforms other active learning methods and random sampling method by showing fast performance improvement.

For example, from iteration step 0 to 1, the performance of LETS substantially increases from 45.5% to 61.6%, while the performances of other methods only increase from 38.3% to around 50.8% in macro-averaged ACSC metric. Other methods achieve a similar performance of 61.6% at iteration step 2-3, which means that LETS can reduce manual labelling effort 2-3 times better with the SemEval dataset. Moreover, it is worth mentioning that LETS achieves significantly (Wilcoxon signed-rank test with $p < .05$) better performances than other methods at the beginning and the end of iteration thanks to the task-specific pre-training and label augmentation. Similar trends are also observed in the micro-averaged ACSC metric. Similar to the result with the Caffeine Challenge dataset, this result shows that the task-specific and the proposed label augmentation can also contribute to better generalisability with the SemEval dataset.

Performance differences between LETS and random sampling method are statistically significant ($p < .05$) throughout entire iteration steps in both micro-and macro-averaged ACSC metrics. Also, performance differences between LETS and other active learning methods are statistically significant ($p < .05$) from iteration 0 to 4 for BALD and from iteration step 0 to 2 for CEAL and least confidence methods, respectively, in both micro and macro-averaged ACSC metrics.

E. DISCUSSION

The proposed LETS integrates multiple components, including task-specific pre-training, label augmentation, and active learning. To investigate the effect of task-specific pre-training with label augmentation separately, we further analyse the performances of a pre-trained model (PT) and task-specific pre-trained model (TSPT) by ablating the label augmentation (LA) component. Fig. 8 and Fig. 9 summarise the ablation study with the Caffeine Challenge dataset and the SemEval dataset, respectively. Note that all models use the proposed active learning method.

From the Fig. 8 and Fig. 9, it is observed that each task-specific pre-training and label augmentation provides performance improvement in the ACSC metrics. Nonetheless, more consistent improvement is observed when both components are applied together. For example, the results from the Caffeine Challenge dataset, as illustrated in Fig. 8, show that task-specific pre-training can contribute to performance improvement and label augmentation can further provide performance boost, especially in early iteration steps.

Similar trends are also observed in the results from the SemEval dataset as illustrated Fig. 9. The major differences are the results from the SemEval dataset are more stable throughout the iteration steps. The results from the Semeval dataset, as illustrated in Fig. 9, show significant differences ($p < .05$) between the task-specific pre-trained model with label augmentation (TSPT+LA) and the pre-trained model (PT) from iteration step 0 to step 4. This suggests that the combination of task-specific pre-training and label augmentation can contribute statistically significant performance improvement for the SemEval dataset, in early iteration steps. Interestingly, each task-specific pre-training and label augmentation can also contribute to the similar performance improvement of combining both of them. This suggests that applying either ask-specific pre-training or label augmentation can be also beneficial for the SemEval dataset.

VI. LIMITATIONS AND FUTURE STUDIES

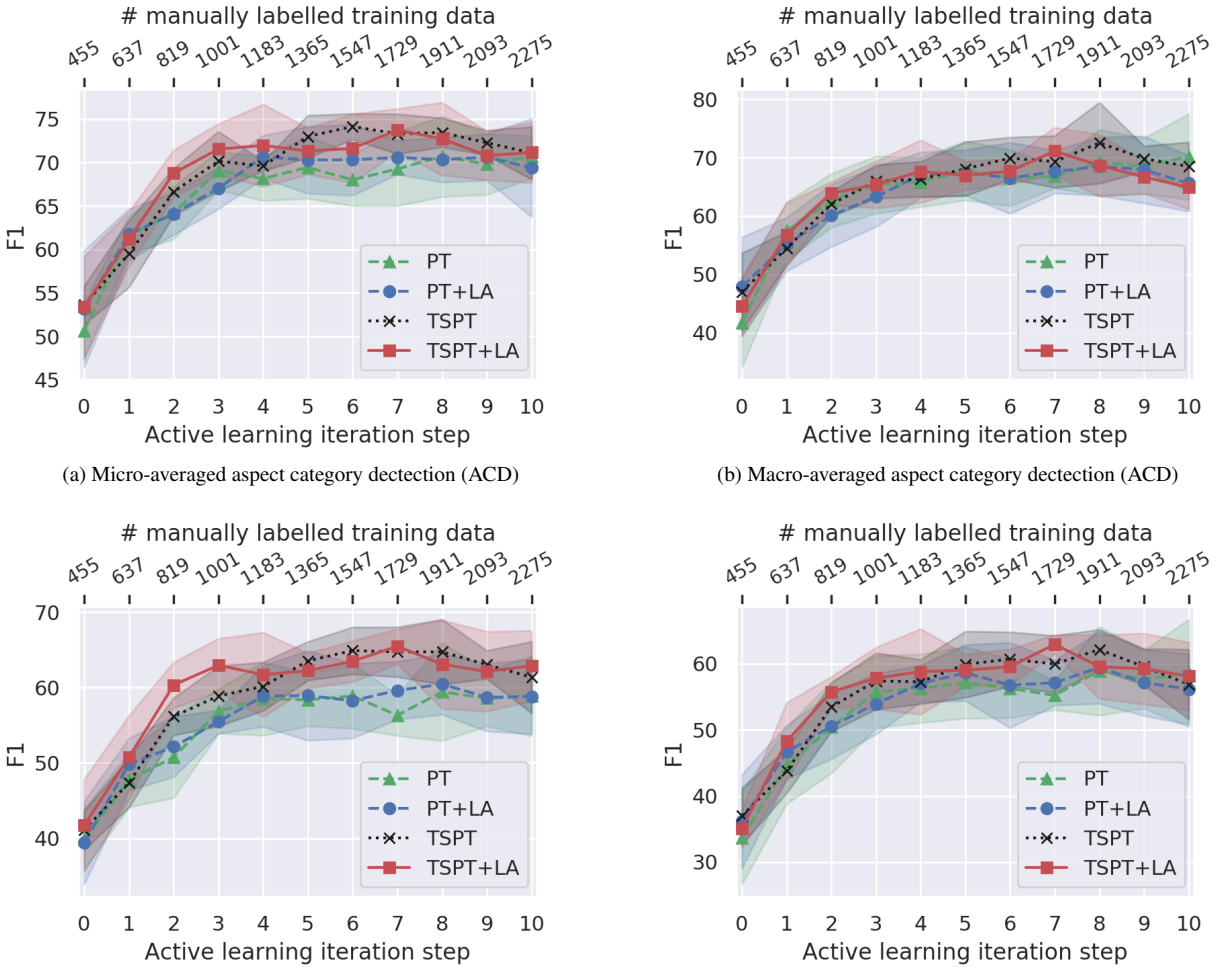
Even though we show the effectiveness of the proposed method by validating with two different datasets, some points can be further studied. Firstly, the Caffeine Challenge dataset is semi-realistic and not collected from actual users of a mobile application. This is mainly because the goal of this paper was to conduct a pilot study of developing an aspect-based sentiment analysis system for the healthcare domain prior to having a mobile application available. Therefore, further study is needed to collect real-world data and conduct experiments to validate the developed system. Since the real-world data are not labelled and the main contribution of this paper is proposing a label-efficient training scheme, we argue that the proposed method can be used to efficiently label the real-world data to further train the system.

The second limitation is the handcrafted rules of the proposed methods. The majority and minority classes were defined based on the frequency in the training sets. Further study could explore an algorithmic approach to distinguish between majority and minority classes. For example, in the active learning setting, minority classes can be dynamically defined based on the labelled data set of the previous iteration step. Also, the proposed label augmentation uses handcrafted dictionaries. A synonym search algorithm by using a lexical database, such as WordNet [38], or a knowledge graph, such as ConceptNet [39], could be used for automatically creating dictionaries for the proposed label augmentation.

Thirdly, a remaining difficulty in applying this work is to know when to start and when to stop active learning iterations. For example, in our experiments (Sec. V), the size of seed data is set to 20% of the training set for the Caffeine Challenge dataset while it is set to 10% of the training set for the SemEval dataset. It is decided based on heuristics and future studies could investigate the optimal size of the seed data. Also, even though the proposed method achieves fast performance improvements at the beginning, it reaches a plateau in the middle of the active learning process. This is because we consider a pool-based active learning scenario, which assumes a large amount of unlabelled data at the beginning of the process and the active learning iteration ends when there is no more data to be labelled. To avoid unnecessary iteration steps, a stopping strategy is needed. Potentially, stopping strategy can be defined based on the stabilisation of predictions [40] or the certainty scores of predictions [41].

VII. CONCLUSION

In this paper, we introduce a new potential application of ABSA applied to health-related program reviews. To achieve this, we collected a new dataset and developed an ABSA



(a) Micro-averaged aspect category dectection (ACD)

(b) Macro-averaged aspect category dectection (ACD)

(c) Micro-averaged aspect category sentiment classification (ACSC)

(d) Macro-averaged aspect category sentiment classification (ACSC)

FIGURE 8. Compared active learning results for ablation study with the Caffeine Challenge dataset. Each line indicates averaged 5-fold results with standard deviation as shade. The bottom X-axis indicates the active learning iteration step and the top x-axis indicates the number of manually labelled training data. Y-axis indicates the performance score. PT and TSTP refer to the model with pre-training and task-specific pre-training, respectively. Masked language modelling is used for task-specific pre-training objective. +LA indicates that label augmentation is applied during the active learning process. All models use the proposed active learning method.

system. Also, we propose a novel label-efficient training scheme to reduce manual labelling efforts. The proposed label-efficient training scheme consists of the following elements: (i) task-specific pre-training to utilise unlabelled task-specific corpus data, (ii) label augmentation to exploits the labelled data, and (iii) active learning to strategically reduce manual labelling.

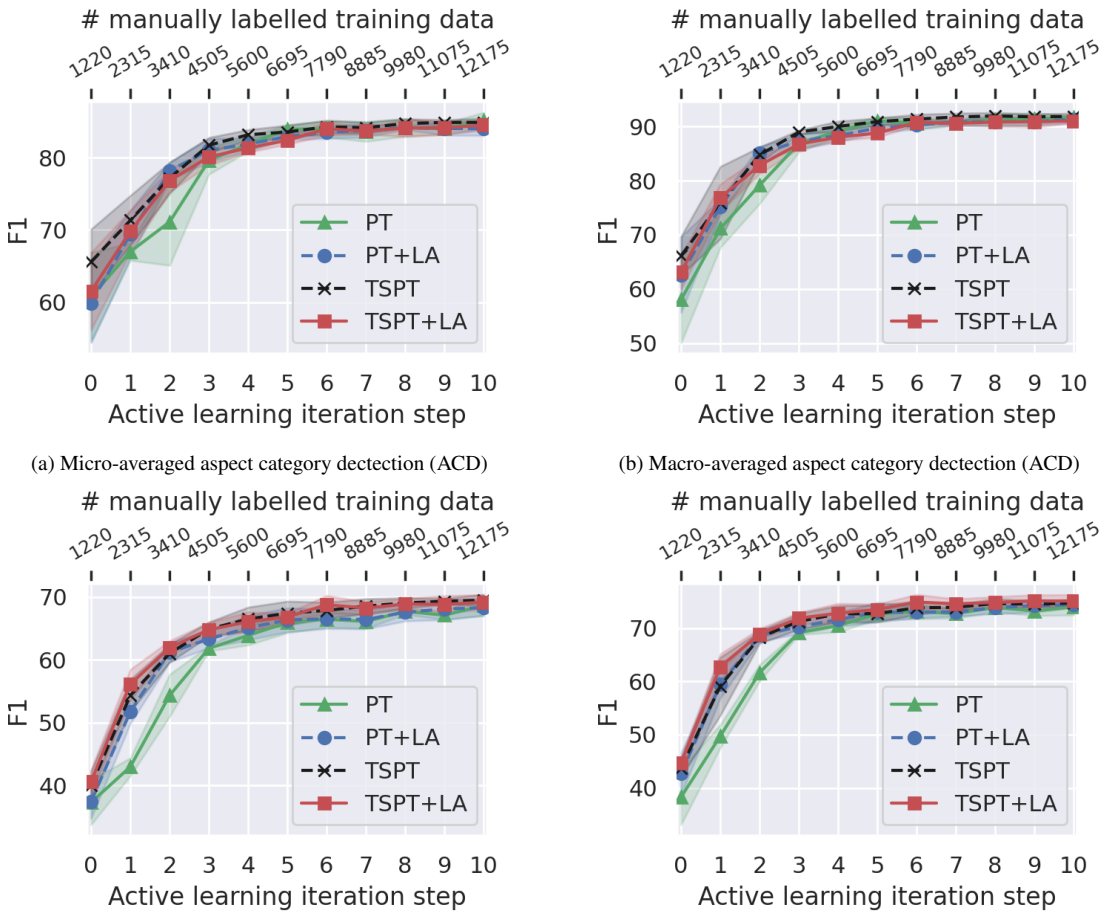
The effectiveness of the proposed method is examined via experiments with two datasets. We experimentally demonstrated the proposed method shows faster performance improvement and achieves better performances over existing active learning methods, especially in terms of the end-to-end evaluation metrics. More specifically, experimental results show that the proposed method can reduce manual labelling effort 2-3 times compared to labelling with random sampling on both datasets. The proposed method also shows better performance improvements than the existing state-of-the-art active learning methods. Furthermore, the proposed method shows better generalisability than other methods thanks to the task-specific pre-training and the proposed label augmentation.

As future work, we expect to collect actual user data from a mobile application and implement the developed ABSA system with the proposed label-efficient training scheme. Moreover, we will investigate a stopping strategy to terminate the active learning process to avoid unnecessary iteration steps.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving



(a) Micro-averaged aspect category dectection (ACD)

(b) Macro-averaged aspect category dectection (ACD)

(c) Micro-averaged aspect category sentiment classification (ACSC)

(d) Macro-averaged aspect category sentiment classification (ACSC)

FIGURE 9. Compared active learning results for ablation study with the SemEval dataset. Each line indicates averaged 5-fold results with standard deviation as shade. The bottom X-axis indicates the active learning iteration step and the top x-axis indicates the number of manually labelled training data. Y-axis indicates the performance score. PT and TSTP refer to the model with pre-training and task-specific pre-training, respectively. Masked language modelling is used for task-specific pre-training objective. +LA indicates that label augmentation is applied during the active learning process. All models use the proposed active learning method.

language understanding by generative pre-training," 2018.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.

[4] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[5] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 208–215.

[6] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[7] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1183–1192.

[8] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017, pp. 252–256.

[9] A. Siddhant and Z. C. Lipton, "Deep bayesian active learning for natural language processing: Results of a large-scale empirical study," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2904–2909.

[10] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2324–2335.

[11] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: a comparative review," *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019.

[12] S. Ruder, P. Ghaffari, and J. G. Breslin, "A hierarchical model of reviews for aspect-based sentiment analysis," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 999–1005.

[13] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 606–615. [Online]. Available: https://www.aclweb.org/anthology/D16-1058

[14] C. Brun and V. Nikoulina, "Aspect based sentiment analysis into the wild," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 116–122.

[15] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 380–385.

[16] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics, Aug. 2014, pp. 27–35. [Online]. Available: https://www.aclweb.org/anthology/S14-2004

[17] W. Xue and T. Li, "Aspect based sentiment analysis with gated convolutional networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2514–2523.

[18] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *China National Conference on Chinese Computational Linguistics*. Springer, 2019, pp. 194–206.

[19] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.

[20] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *SIGIR'94*. Springer, 1994, pp. 3–12.

[21] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.

[22] A. Shelmanov, V. Liventsev, D. Kireev, N. Khromov, A. Panchenko, I. Fedulova, and D. V. Dylov, "Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 482–489.

[23] L. E. Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, "Active learning for bert: An empirical study," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7949–7962.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017, pp. 6402–6413.

[25] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9368–9377.

[26] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–35, 2020.

[27] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 35–50.

[28] J. Gonsior, M. Thiele, and W. Lehner, "Weakal: Combining active learning and weak supervision," in *International Conference on Discovery Science*. Springer, 2020, pp. 34–49.

[29] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[30] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[32] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.

[33] D. Reker, "Practical considerations for active machine learning in drug discovery," *Drug Discovery Today: Technologies*, 2020.

[34] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7935–7948.

[35] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 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. 6382–6388. [Online]. Available: https://www.aclweb.org/anthology/D19-1670

[36] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 127–136.

[37] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[38] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[39] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[40] M. Bloodgood and K. Vijay-Shanker, "A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 2009, pp. 39–47.

[41] J. Zhu, H. Wang, E. Hovy, and M. Ma, "Confidence-based stopping criteria for active learning for data annotation," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 6, no. 3, pp. 1–24, 2010.

[42] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, pp. arXiv–1910, 2019.

.

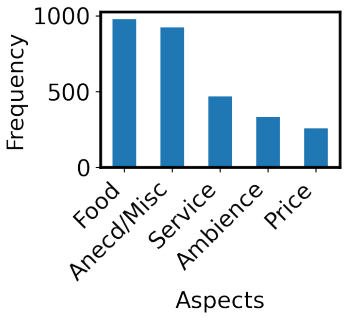


FIGURE 10. Aspect category distribution of the training set from the SemEval dataset. Anecd/Misc refers Anecdotes/Miscellaneous aspect category.

APPENDIX A EXAMPLES OF THE COLLECTED DATA

Table 5 shows examples of the collected data used for experiments.

APPENDIX B EXPLANATION OF ASPECT CATEGORIES

Table 6 summarises the explanation and examples of aspect categories used in the paper.

APPENDIX C ASPECT CATEGORY DISTRIBUTION OF THE SEMEVAL DATASET

Fig. 10 illustrates the aspect category distribution of the training set from the SemEval dataset used for the experiments. As it is shown in the figure, the SemEval dataset is imbalanced and we define {Food, Anecdotes/Miscellaneous} and {Service, Ambience, Price} as majority and minority aspect categories, respectively.

Imagine you **successfully finished** the challenge.
Q1: How was your experience with this challenge and why?
Answer (pos): My experience was great. I felt that my experience was personalized and I really was able to fall asleep faster and stay asleep longer by giving up caffeine after 1pm. It was a lot easier than expected.
Answer (neu): It was okay. While I did find it helpful to give up caffeine after 1pm to help with my sleep, it was difficult for me to give up and almost felt as if I were detoxing from caffeine.
Answer (neg): My experience was not very good. While I was able to give up caffeine after 1pm, it gave me a headache as I must have been going through withdrawals and in turn, these headaches kept me up later than I would have wanted.
Q2: Could you tell me how reducing caffeine affected you?
Answer (pos): Reducing caffeine really affected me positively. I was easily able to give the caffeine up after 1pm and in turn, I fell asleep much faster and didn't wake up throughout the night as I normally would.
Answer (neu): It was an okay experience. While I slept better, it was difficult for me to give up the caffeine, especially chocolate when I crave a snack after work.
Answer (neg): My experience was not very good. Because I gave up caffeine, I think my wellbeing was negatively affected because I then had a headache which made getting to sleep difficult. I think I actually lost sleep due to this.
Imagine you **was not able to complete** the challenge.
Q1: How was your experience with this challenge and why?
Answer (pos): While it was difficult for me to give up my afternoon and evening caffeine so I could not complete the challenge, I still had a positive experience as I did sleep better on the nights that I did successfully complete the challenge for the day.
Answer (neu): The experience was just okay for me. Because I did not successfully complete the challenge, I am not sure that I saw all of the benefits. I would like to try again in the future.
Answer (neg): I didn't like having to give up the caffeine. I kept getting headaches and for that reason I went back to the caffeine and did not successfully complete the challenge.
Q2: Could you tell me how reducing caffeine affected you?
Answer (pos): Reducing caffeine affected me by allowing me to go to sleep earlier and stay asleep longer. Therefore, I felt better and more refreshed when I woke up in the morning.
Answer (neu): It affected me in an okay way. While my sleep did tend to be better, I struggled with actually giving up the caffeine. This is something I would have to work at.
Answer (neg): It affected me negatively because while I was giving up the caffeine, I actually saw an increase in headaches and because of this, I also saw a lack of sleep.

TABLE 5. Example of question and answers. This example shows 12 different responses from a single participant.

APPENDIX D IMPLEMENTATION AND TRAINING SETTINGS

All experiments were performed on the Windows 10 operating system and the detailed specification of hardware and software is summarised in Table 7. For model implementation, PyTorch version of BERT with the pre-trained weights (`bert-base-uncased`) [42] was used as the pre-trained model (PT). During task-specific pre-training, the pre-trained model is further trained on the end task corpus. For task-specific pre-training, we adopt masked language modelling [1] with masking probablity $p = 0.15$. During task-specific pre-training, randomly sampled 10% of training data is used as a validation set for early-stopping.

For fine-tuning, 5-fold cross validation splits are created by using K-Folds cross-validator function from scikit-learn library[6]. Also, a final dense layer with softmax function is added and cross entropy loss is used. Since the focus of this paper is active learning experiments, we did not conduct hyperparameter tuning experiments but used hyperparameter values based on the recent study [18] as summaries in Table 8.

Aspect	Explanation/Examples
Sleep quality	Impact on sleep quality. Positive: Sleep quality has improved. Negative: Sleep quality has been worsened.
Mood	Experience related to mental state. Positive: became calm or relaxed. Negative – felt nervous/anxious, experienced mental drain, or negative thoughts.
Energy	Impact on energy and concentration. Positive: Had/felt ore energy during day. Negative: Tired during the day or couldn't concentrate at work.
Missing caffeine	Feeling of caffeine deprivation. Positive: Did not miss caffeine products. Negative: Missed the taste of caffeine or didn't like decaffeinated alternatives.
Difficulty level	Difficulty of the challenge. Positive: Challenge was easy/easier than thought. Negative: Too difficult to change the habit.
Physical withdrawal symptoms	Impact on physical state. Positive: Physical state has improved. Negative: Experienced headache, stomach aches, or any other physical withdrawal symptoms.
App experience	Experience with app. Positive: App was supportive or reminder/recommender was helpful. Negative: User experience of app was bad or the reminder was annoying.

TABLE 6. Explanation and examples of aspect categories.

Item	Specification
CPU	Intel®Xeon®W-2123 CPU @ 3.60 GHz
GPU	NVIDIA GeForce GTX 1080 ti, 11 GB memory
Graphic driver	NVIDIA graphic driver version 416.34
CUDA	Version 10.0
OS	Windows 10, 64-bit
Python	Version 3.6.6
Pytorch	Version 1.5.1

TABLE 7. Detailed implementation specification.

Hyperparameter	Assignment
training epoch	4
batch size	32
learning rate	$2e-5$
drop out	0.1
optimaser	AdamW
training epoch	4
batch size	32
learning rate	$2e-5$
drop out	0.1
optimaser	AdamW
classificaiton layer	feedforward

TABLE 8. Hyperparameters for task-specific pre-training (top) and fine-tuning (bottom).

[6]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

APPENDIX E PRE-DEFINED DICTIONARIES FOR LABEL AUGMENTATION

Pre-defined dictionaries were used for label augmentation. For the Caffeine Challenge dataset, the list of minority aspect categories and the list of similar words for each aspect categories are defined as:

- Missing caffeine: [Missing caffeine, Dislike decaffeine, Need caffeine, Caffeine addiction]
- Difficulty level: [Difficulty level, Hard to finish, cannot complete, Too difficult]
- Physical withdrawal symptoms: [Physical withdrawal symptoms, Headache, Pain, Jitter]
- App experience: [App experience, UI, UX, Design]

For SemEval dataset, the list of minority aspect categories and the list of similar words for each aspect categories are defined as:

- Service: [Service, Staff][7]
- Ambience: [Ambience, Atmosphere, Decor]
- Price: [Price, Bill, Quality[8]]



HEEREEN SHIM received an MSc in Electrical Engineering from Chung-Ang University, South Korea in 2018. Currently, she is pursuing her Ph.D. in Engineering Technology at KU Leuven, Belgium within a research framework HEART (HEalth-related Activity Recognition system based on IoT) funded by European Commission. She is also working at Philips Research since 2018 where she has been working on natural language processing for health-related behaviour analytics based on free-text. Her research interests lie in scalable machine learning algorithms and natural language processing for a conversational agent particularly for the healthcare domain.



DIETWIG LOWET has obtained his Master of Science degree in Physics at the University of Antwerp on 1996. Afterwards, he obtained a Professional Doctorate in Engineering (PDEng) at the Computer Science Faculty of the Eindhoven University of Technology. He has been working at Philips Research since 1999 where he has worked on embedded software architectures for (connected) consumer devices, data analytics and insight generation from personal health data for digital lifestyle coaching services. His current work is on natural language processing and understanding to analyse health-related free text input.

[7]During experiments, we observed that adding more labels for Service aspect category harms the performance.

[8]Quality is not a similar word for price but it is used because the training data set contains reviews mentioning price-quality relationship.



STIJN LUCA holds an MSc in Mathematics from KU Leuven (Belgium, 2003) and a PhD in mathematics from Hasselt University (Belgium, 2007). Currently, he is an assistant professor at the Department of Data Analysis and Mathematical Modelling of the Faculty of Bioscience Engineering at Ghent University (Belgium). Previously, he was a postdoctoral fellow at KU Leuven (Belgium) and a visiting scholar at the University of Oxford (2014–2015). Luca's research interests surround the theory and methods of statistical data analysis and their applications in life sciences.



BART VANRUMSTE received an MSc in Electrical Engineering and MSc in Biomedical Engineering both from Ghent University in 1994 and 1998, respectively. In 2001, he received a Ph.D. in Engineering from the same institute with the title: 'EEG Dipole Source Analysis in a Realistic Head Model'. He worked as a post-doctoral fellow from 2001 until 2003 at the Electrical Computer Engineering Department of the University of Canterbury, New Zealand. From 2003 until 2005, he was a post-doctoral fellow at the Department of Electrical Engineering (ESAT) in the STADIUS division at KU Leuven. In 2005, he was appointed faculty member initially at the University of Applied Sciences Thomas More and since 2013 in the Faculty of Engineering Technology of KU Leuven. Bart Vanrumste currently teaches courses in Digital Signal Processing, Digital Image Processing, and Machine Learning. He is a member of the eMedia research lab at Group T and a member of the ESAT-STADIUS division. His research interests are decision support in healthcare in general and ICT applications for aging in place in particular. His current research activities focus among others on multimodal sensor integration for monitoring of older persons and patients with chronic diseases. He is a senior member of IEEE Engineering in Medicine and Biology Society and member of the International Society for Bioelectromagnetism.