# One for "All": A unified model for fine-grained sentiment analysis under three tasks

Heng-yang Lu[1,2], Jun Yang[3,*], Cong Hu[1], Wei Fang[1]

[1] Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, Jiangsu, China.
[2] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu, China.
[3] Marcpoint Co., Ltd., Shanghai, China.

Corresponding Author:
Jun Yang
No. 456, Hongcao Road, Xuhui District, Shanghai, 200233, China
Email address: jun.yang@marcpoint.com

## Abstract

**Background.** Fine-grained sentiment analysis is used to interpret consumers' sentiments, from their written comments, towards specific entities on specific aspects. Previous researchers have introduced three main tasks in this field (ABSA, TABSA, MEABSA), covering all kinds of social media data (e.g. review specific, questions and answers, and community-based). In this paper, we identify and address two common challenges encountered in these three tasks, including the low-resource problem and the sentiment polarity bias.

**Methods.** We propose a unified model called PEA by integrating data augmentation methodology with the pre-trained language model, which is suitable for all the ABSA, TABSA and MEABSA tasks. Two data augmentation methods, which are entity replacement and dual noise injection, are introduced to solve both challenges at the same time. An ensemble method is also introduced to incorporate the results of the basic RNNs-based and BERT-based models.

**Results.** PEA shows significant improvements on all three fine-grained sentiment analysis tasks when compared with state-of-the-art models. It also achieves comparable results with what the baseline models obtain while using only 20% of their training data, which demonstrates its extraordinary performance under extreme low-resource conditions.

## Introduction

Consumers worldwide have posted trillions of text comments on online shopping sites and social platforms to express their opinions. The efficiency of how modern merchandisers drive insights from those opinions would be the key to their success in the data-driven era. Sentiment analysis is such a solution for businesses to understand consumers' opinions effectively. Traditional coarse-grained sentiment analysis aims to identify the sentiment polarity of the given sentence. Different from that, fine-grained sentiment analysis is managed to match sentiments with corresponding entities and aspects in the given sentence. For example, given the comment "I've used MacBookPro, it's convenient." Coarse-grained sentiment analysis describes the whole sentence a positive sentiment. Fine-grained sentiment analysis describes a positive sentiment towards MacBookPro (entity) on its convenience level (aspect), which is a provided (sentence, aspect, entity) pair. Previous researchers have introduced three tasks on fine-grained sentiment analysis towards entities and aspects (definitions and 2 examples are illustrated in Table 1) :

1.  Aspect-Based Sentiment Analysis (**ABSA**),
2.  Targeted Aspect-Based Sentiment Analysis (**TABSA**),
3.  Multi-Entity Aspect-Based Sentiment Analysis (**MEABSA**).

ABSA was primarily based on the review-specific data acquired from E-commerce or life service websites (e.g. Amazon, Yelp) where there is only one or even no entity mentioned in the data. Although performing well on consumer reviews, models designed for ABSA have limited performance on posts coming from social platforms (e.g. Twitter, Reddit) where there are multiple entities and aspects mentioned. For example, a software engineer on Twitter wrote "I've used MacBookPro, it's convenient. But now I switched to ThinkPad because it's just as convenient and has a better price." There are two entities introduced: MacBookPro and

82  ThinkPad. For each of the entities, sentiments on the convenience level (aspect 1) are the same
83  while sentiments on the price (aspect 2) are different. TABSA was proposed by (Saeidi et al.,
84  2016) to handle such multi-entity and multi-aspect cases. This task was based on the *SentiHood*
85  dataset acquired from the question answering platform, which involves 2 entities of the same
86  kind (e.g. tourist attractions) and 15 aspects. However, in reality, not only do consumers compare
87  entities of that same kind but also should they talk about multi-kind entities. Yang (Yang et al.,
88  2018)  proposed MEABSA with the *BabyCare* dataset acquired from a community-based
89  platform. It involves hundreds of multi-kind entities (e.g. powdered milk, diapers, and infant
90  medicines) and hundreds of aspects. The increase in the number of entities and aspects makes
91  MEABSA the most challenging task among the three.
92  Most previous works are designed for only one of the tasks, it is more practical to design a
93  unified model, which is available for all three tasks. What's more, the Recurrent Neural Network
94  (RNN)-based models (Yang et al., 2018; Yang et al., 2019; Xu et al., 2020) and BERT-based
95  models (Sun, Huang & Qiu, 2019) are two kinds of recently proposed basic models for fine-
96  grained sentiment analysis, which have shown effectiveness. The RNN-based models have the
97  advantages of considering the global sequence, and the BERT-based models are good at
98  considering local attention. It is promising to improve the predictions of sentiments by making
99  use of both advantages.
100 Additionally, there are two main challenges encountered in the ABSA, TABSA, and MEABSA
101 tasks. The first challenge is the **low-resource problem**, also known as the insufficient data
102 problem. This is often caused by the large time and money required by manual annotation. The
103 low-resource problem is even more prevalent in sentiment prediction towards entities and aspects
104 due to the increasing complexity of data annotation: for example, if there are 3 entities and 2
105 aspects mentioned in the text, one needs to annotate 6 (3*2) instances for each of the entity
106 aspect combinations. This explains the fact that 59% of the entity aspect combinations are
107 annotated 5 times or less in the *BabyCare* dataset. The second challenge is the **polarity bias**
108 **problem**. It reduces task performance when entities' sentiment polarity distribution is not
109 uniform in the training set. For example, if an entity is mostly labeled positive in the training set,
110 it will be more likely to be predicted positive regardless of the context. This problem is mainly
111 caused by the inconsistent polarity distributions between the training set and test set from the
112 perspective of entities.
113 This paper aims to propose a unified model for fine-grained sentiment analysis, which is
114 available for ABSA, TABSA and MEABSA tasks. The main contributions of this paper include:
115 •   To the best of our knowledge, this is the first work unifying the ABSA, TABSA, and
116     MEABSA tasks together, providing an all-in-one solution to fine-grained sentiment analysis.
117 •   We propose a unified model, which combines both advantages of RNN-based models and
118     BERT-based models with ensemble methods. This model achieves outstanding performance
119     in all the ABSA, TABSA, and MEABSA tasks.
120 •   This paper considers the low-resource and polarity bias problems in the fine-grained
121     sentiment analysis for the first time. Two data augmentation methods include entity
122     replacement and noise injection are designed to deal with the problems.

123

## Literature Reviews

**Research on fine-grained sentiment analysis**

There are abundant researches on the ABSA task. LSTM (Tang et al., 2016) and an attention mechanism (Wang et al., 2016) have been applied to deal with the ABSA task in early time. Following works include applying memory network-based (Tang, Qin & Liu, 2016) and attention-based (Chen et al., 2017) method to LSTM models, involving two stacked LSTMs (Xu et al., 2020) and so on. More recent models such as capsule network (Chen & Qian, 2019; Du et al., 2019), graph convolutional network model (Zhang, Li & Song, 2019), graph attention network (Wang et al., 2020), bi-level interactive graph convolution network (Zhang and Qian, 2020) are also used for ABSA task. Zhu and associates have exploited the interaction between the aspect category and the contents under the guidance of both sentiment polarity and predefined categories, and the proposed aspect aware learning framework has achieved satisfying performance in ABSA (Zhu et al., 2019). The interactive relationships among aspect term extraction, opinion term extraction, and aspect-level sentiment classification have been investigated to encode collaborative signals for unified ABSA tasks (Chen and Qian, 2020). The pre-trained model such as RoBERTa has also been applied to improve ABSA with induced trees (Dai et al., 2021).

Saeidi and associates (Saeidi et al., 2016) first proposed the TABSA task with SentiHood dataset. Following works include using additional commonsense knowledge (Ma, Peng & Cambria, 2018), developing a delayed memory update mechanism (Liu, Cohn & Baldwin, 2018), extending LSTM by adding the external knowledge (Khine et al., 2019) and so on. Additionally, Ye and Li proposed a recurrent entity memory network with word-level information and sentence-level hidden memory for TABSA (Ye and Li, 2020). In recent years, pre-trained language model is also applied to capture the dependence on both targets and aspects for sentiment prediction (Wan et al., 2020). BERT model has been applied to TABSA task. For example, auxiliary sentence has been found useful in TABSA when BERT model is applied (Sun, Huang & Qiu, 2019). Similarly, Hong and associates (Hong & Song, 2020) further fine-tune the pre-trained BERT model on SentiHood dataset. What's more, a context-guided softmax-attention and context-guided quasi-attention method is proposed to perform aspect categorization and TABSA at the same time (Wu and Ong, 2020).

Yang and associates (Yang et al., 2018) first proposed the MEABSA task and contributed a dataset named *BabyCare*. They also proposed the Context memory, Entity memory and Aspect memory model (CEA) with RNN and deep memory networks. To improve the performance on long and complex text, an extended model of combining dependency trees with deep neural networks was proposed (Yang et al., 2019). The data sparsity challenge, also known as the cold-start problem, has also been investigated in MEABSA, which designed the frequency-guided attention mechanism to solve the problem (Song et al.,2019).

**Research on data augmentation in NLP**

162  To alleviate the low-resource problem in various NLP tasks, data augmentations have been
163  applied in previous works. The optional strategies mainly include word replacement, noise
164  injection, text generation and so on. For example, it is useful to generate additional training
165  examples that contain rare words in synthetically created contexts for machine translation
166  (Fadaee, Bisazza & Monz, 2017).  Another similar idea injected low-resource words into high-
167  resource sentences to improve the low-resource translation task (Xia et al., 2019). Additionally,
168  data augmentations such as synonym replacement and delexicalization have been applied to the
169  NER task (Dai & Adele,  2020) and dialogue language understanding (Hou et al., 2018)
170  respectively. Kim and associates (Kim, Roh & Kim, 2019) proposed a method for spoken
171  language understanding by introducing noise in all slots without classifying types of slots to
172  improve the performance of low-resource dataset with "open-vocabulary" slots.
173  **Research on Bias problems in NLP**
174  Bias, such as racial bias and gender bias (Kiritchenko & Mohammad, 2018; Thelwall, 2018), is
175  also a trending topic of concern in different NLP researches. For example, Zhao and associates
176  (Zhao et al., 2018) tried to mitigate gender bias by creating an augmented dataset identical to the
177  original one by replacing the entities such as "he" or "she".  Another work formally proposed the
178  Counterfactual data augmentation (CDA) for gender bias mitigation in the coreference resolution
179  task, by replacing every occurrence of a gendered word in the original corpus with its flipped one
180  (Lu et al., 2020).
181  Recently, there are some related works to deal with the low-resource and polarity bias problems
182  in coarse-grained sentiment analysis, which aims to predict the sentiments of the given posts. An
183  early work introduced a bias-aware thresholding method motivated by cost-sensitive learning
184  (Iqbal, Karim & Kamiran, 2015). Recent works include designing a sentiment bias processing
185  strategy for the lexicon-based sentiment analysis (Han et al., 2018), and using the generation-
186  based data augmentation method to deal with the low-resource problem in coarse-grained
187  sentiment analysis (Gupta, 2019). To the best of our knowledge, there is no recent work
188  discussing solutions to low-resource or polarity bias problems in fine-grained sentiment analysis.
189
190  # Methods
191  ABSA, TABSA and MEABSA are three widely discussed tasks for fine-grained sentiment
192  analysis, whose common objective is to predict the sentiment towards each aspect of each target
193  entity. The detailed comparisons and examples can be found in Table 1 in the introduction
194  section. This section introduces the methodologies, which we used to unify the ABSA, TABSA,
195  and MEABSA tasks together with the same architecture. The proposed all-in-one solution to
196  **P**redict sentiment towards **E**ntities and **A**spects is named **PEA**. Fig. 1 demonstrates the graphical
197  abstract of the PEA model.
198
199  Firstly, the unified problem setting of fine-grained sentiment analysis covering ABSA, TABSA
200  and MEABSA is as follows.
201  **Problem Setting**

202    Given a post $Post_m = [w_1, w_2, ..., w_T]$, with an entity set (if available) $E_m = \{entity_1, entity_2$

203    $, ..., entity_{|E_m|}\}$ and an aspect set $A_m = \{aspect_1, aspect_2, ..., aspect_{|A_m|}\}$. For the words or

204    multiple words in $Post_m$, which are corresponding to the entities or aspects in $E_m$ or $A_m$, we call

205    them entity terms and aspect terms. The fine-grained sentiment analysis aims to predict the

206    sentiment $y_{entity_i}^{aspect_j}$ towards the given $aspect_j$ of the certain $entity_i$ in $P_m$.

207    For the ABSA task, the entity set $E_m = \emptyset$ and the prediction target is simplified to $y^{aspect_j}$.

208    For the TABSA task, in each post $Post_m$, there is only one or two entities in the entity set, where

209    $|E_m| = 1$ or $|E_m| = 2$. The prediction target becomes $y_{entity_i}^{aspect_j}$ towards all the aspects for the

210    target entity in $Post_m$.

211    For MEABSA, the most challenging task, there are multiple entities and aspects in $Post_m$, where

212    $|E_m| \geq 1$ and $|A_m| \geq 1$. It aims to predict $y_{entity_i}^{aspect_j}$ towards the mentioned aspects for every entity

213    $entity_i$ in $Post_m$.

214    The general training workflow of PEA includes:

215    (1) given an original training set $D$, generate a new training set $D'$ based on entity replacement.

216    For the ABSA task, there is no entity involved, so the entity replacement step is skipped and $D'$

217    $= D$. For TABSA and MEABSA, entity replacement is conducted to get an entity-replaced

218    dataset $PD$, and $D' = D \cup PD$. The entity replacement used in PEA is introduced in the first part

219    of subsection "Data Augmentation".

220    (2) An RNN-based model is trained on the new training set $D'$ as one of the basic models. The

221    dual noise injection is conducted on the input posts, entities and aspects to get the noise-injected

222    vectors. The dual injection used in PEA is introduced in the second part of subsection "Data

223    Augmentation". Then, we take an attentional recurrent neural network-based model, CEA (Yang

224    et al., 2018), as an example, to be the basic model, whose output is the predicted sentiment

225    polarity distribution of the given inputs. It is introduced in the first part of subsection "Basic

226    Models".

227    (3) A pre-trained language model is trained on the new training set $D'$ as the other basic model.

228    Auxiliary question sentences are constructed for training the BERT-based model, which can

229    predict fine-grained sentiment polarity distribution with the given inputs. The detailed design is

230    described in the second part of subsection "Basic Models".

231    (4) Finally, the ensemble method is applied to fuse the predicted sentiment polarity distribution

232    of the RNN-based and BERT-based model as the outputs of PEA, which is the final predicted

233    sentiment polarity. The fusion strategy is introduced in the third part of subsection "Fusion

234    Strategy".

235    **Data Augmentation**

236  Data augmentation is widely used to improve learning performance, prevent overfitting, and
237  increase robustness under low-resource conditions. This section illustrates two innovative, task-
238  specific data augmentation methods that are deployed in the model.
239  **Entity Replacement.** The low-resource problem in fine-grained sentiment analysis mainly
240  comes from entities in the posts. This problem can be alleviated by increasing the low-resource
241  entities. Among the data augmentation methods used in recent works for alleviating the low-
242  resource problem in other NLP tasks, replacing words in context with similar ones is a viable
243  data augmentation method (Fadaee, Bisazza & Monz, 2017; Xian et al., 2019; Dai & Adel,
244  2020). Usually, similar words can be extracted from word similarity calculation (Wang &Yang,
245  2015), and can also be extracted from a handcraft ontology such as WordNet.
246  In previous works, any word in a sentence can be replaced. This kind of replacement is
247  extremely risky in fine-grained sentiment analysis tasks. For example, if a sentiment word, such
248  as "happy", was replaced, it would unintentionally change the sentiment polarity at the same
249  time. To avoid this kind of situation, we proposed the entity replacement method which
250  successfully addresses this problem. Entity replacement is used to generate pseudo instances for
251  training. The entire process involves 3 steps:
252  • Creating a duplicate of the original training set $D$.
253  • Replacing each entity in the duplicated dataset with the target entity to get an entity-replaced
254  dataset $PD$.
255  • Combining the original dataset with the entity-replaced dataset as the new training dataset $D^{'}$
256  $= D \cup PD$ to train models.
257  In step 2, target entities are selected dynamically based on the scarcity of entities in the original
258  training set so that every entity will have sufficient training instances eventually. In other words,
259  the fewer times an entity presents in the original training set, the more likely it will be selected as
260  the target entity. The detailed probability that an entity is selected is calculated as follows:

261
$$\mathrm{P}(entity_i) = \frac{|mention(entity_i)|^{-1}}{\sum_{j=1}^{|E|}|mention(entity_j)|^{-1}}, \forall\, i \in [1, |E|] \tag{1}$$

262  where $E = \bigcup_{E_m \in D} E_m$ is the total entity set in the original training dataset $D$,

263  $|mention(entity_i)|$ represents the number of instances mentioning $entity_i$ in $D$. $x^{-1}$ is an

264  inverse proportional function, where $x^{-1} = \frac{1}{x}$.

265  Table 2 shows an example of such a replacement. Besides increasing the number of training
266  instances, we think data augmentation also helps solve the polarity bias problem. For example, if
267  an entity is always labeled positive in the training set, it will be more likely to be predicted
268  positive no matter what the post is about. The proposed data augmentation methods help the
269  polarity balance for entities because the entity may be replaced into a positive or neutral or
270  negative expression randomly.
271  To conclude, the low-resource entity replacement is designed to increase the number of training
272  instances, especially for the low-resource entities, and help solve the polarity bias problem in
273  sentiment prediction towards multiple entity settings.

274     **Dual Noise Injection.** To improve the generalization ability of PEA, we also involve the noise
275     injection method. In previous NLP tasks, such as machine translation (Cheng et al., 2018) and
276     spoken language understanding (Kim, Roh & Kim, 2019), it has shown the effectiveness of
277     improving the model's generalization ability by injecting noises. In these works, noise is usually
278     injected into the context representation for the post directly. For fine-grained sentiment analysis,
279     the inputs include context texts, entities, entity terms, aspects and aspect terms. It is not
280     applicable to only inject noises on context representations like previous works. Therefore, we
281     propose the idea of dual noise injection: a noise is injected into the representation of entity and
282     entity terms in the context at the same time. A similar practice is performed on the aspect and
283     aspect terms.
284     In this task, the dual noise injection is used to simulate new entities and new aspects, enabling
285     the model to make better predictions when it comes across low-resource entities or aspects.
286     Following the common choice of previous works (Cheng et al., 2018; Kim, Roh & Kim, 2019),
287     we also use the Gaussian noise to inject noises into the embedding space of posts, entities and
288     aspects. Fig. 2 is an example to illustrate the detailed processes of dual noise injection.
289     The dual noise injection consists of 3 steps:
290     •    We first express the post, entity, and aspect in vectors space $v_w \in \mathbb{R}^{T \times k}$, $v_e \in \mathbb{R}^k$, $v_a \in \mathbb{R}^k$,
291        where $v_w = \{v_{w_1}, ..., v_{w_T}\}$, $T$ represents the number of words in the post, and $k$ is the
292        dimension of representations. The embedding vectors can be initialized by GloVe
293        (Pennington, Socher & Manning, 2014).
294     •    Then we sample noise vectors $n_e \in \mathbb{R}^k$ and $n_a \in \mathbb{R}^k$ for entity and aspect respectively from
295        the Gaussian distribution.
296     •    At last, we extract indicator vector $i_e = \{i_e^0, ..., i_e^T\}$ for entity terms indicating the location of
297        entity terms in the post. Each element in $i_e$ is binary. $i_e^t$ is set to 1 when the $t^{th}$ word in the
298        post is an entity term, otherwise, it is set to 0. Note that an entity term may consist of one or
299        more words. In the same manner, we can get an indicator vector $i_a$ for aspect term. Then, we
300        inject the noise to the entity, the aspect, and the post:

$$v_e' = v_e + n_e. \tag{2}$$

$$v_a' = v_a + n_a. \tag{3}$$

$$v_{w_i}' = v_{w_i} + i^e \times n_e + i^a \times n_a. \tag{4}$$

304     In step 2, the same noise vector (e.g. $n_e$) needs to be applied to the entity and entity term. This is
305     to ensure the new-generated entity and entity term remain the same relative location in the
306     embedding space. We also apply the same noise vector (e.g. $n_a$) to the aspect and the aspect term
307     in the same manner. The noise injected into the entity and aspect does not have to be equal.
308     Also, if the noise level is not large enough, it won't substantially change the effect of injections.
309     In order to test what is the best noise level in this case, we conduct experiments to determine the
310     settings, which is introduced in section "Experimental Settings".
311     **Basic Models**

312   Recently, both RNN-based models and BERT-based models have shown effectiveness in the
313   fine-grained sentiment analysis (Yang et al., 2018; Yang et al., 2019; Sun, Huang & Qiu, 2019;
314   Xu et al., 2020;). Due to the different structures of RNN and BERT, both kinds of models have
315   advantages and weaknesses respectively. PEA incorporates both models to help make the final
316   prediction more accurate.
317   **RNN-based Model for Fine-grained Sentiment Analysis.** The CEA model is designed for
318   MEABSA task, and can also be used for ABSA and TABSA tasks. It takes the word vectors of
319   the post, the entity vectors and aspect vectors as inputs, and predicts the fine-grained sentiments
320   towards the given aspect of the entity. To incorporate noise injection with CEA, we feed the
321   noise-injected vectors to CEA, the general structure of noise-injected CEA is as Fig. 3 shows.
322   Firstly, we feed every noise-injected word vector $v_{w_i}'$ in the post to CEA. An LSTM layer is

323   applied to extract the semantics of the post after a few data processing layers. After that, a deep
324   memory network is applied to update entity and aspect representations with the given noise-
325   injected entity vector $v_e'$ and aspect vector $v_a'$. The updated representations are fed into a dense
326   layer to predict the final sentiment. For detailed explanation of CEA, refer to the original paper
327   (Yang et al., 2018).
328   Because CEA requires entities and aspects as inputs, it is naturally suitable for the TABSA and
329   MEABSA tasks. For the ABSA task, if there is no entity mentioned in the post, we can set the
330   entity vector to a zero vector as the input. This makes the CEA-based basic model be able to deal
331   with all the ABSA, TABSA and MEABSA tasks.
332   **Pre-trained Language Model for Fine-grained Sentiment Analysis.** The pre-trained language
333   model is useful for enabling low-resource tasks to benefit from a huge amount of unlabeled data
334   by pre-training. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al.,
335   2018) is one of the key innovations in language representation learning (Howard & Ruder, 2018;
336   Peters et al., 2018). It has achieved good results in many natural language processing tasks
337   (Acheampong et al., 2021; van Aken et al., 2019).
338   BERT uses bidirectional pre-training for language representations, and it is pre-trained on two
339   tasks: *masked language model* for understanding the relationship between words, and *next*
340   *sentence prediction* for understanding the relationship between sentences for downstream tasks.
341   The design of pre-training makes use of a huge amount of unlabeled data, making it suitable for
342   low-resource situations. Thus, we incorporate BERT to further enhance performance.
343   Sun (Sun, Huang & Qiu, 2019) argued that constructing an auxiliary question sentence for the
344   BERT model is useful in the TABSA task. We follow the conclusion and make the auxiliary
345   question sentence for the entity and aspect with the template of "What is the sentiment towards
346   the [aspect] of [entity]?". Then the sentiment classification task is turned into a sentence pair
347   classification task. The label set of this setting includes {Positive, Neutral, Negative}. The BERT
348   model takes two paragraphs as input with the token [CLS] at the beginning and [SEP] at the end
349   of each paragraph. We set the post as the first paragraph and the auxiliary question sentence as
350   the second. Here is an example.

|  |  |
|---|---|
| **Input:** | [CLS] I've used MacBookPro, it's convenient. [SEP] What is the |

> sentiment towards the convenience of MacBookPro? [SEP]
> **Output:**    Positive

351   By constructing auxiliary question sentences along with the posts, we can generate inputs
352   suitable for training BERT-based models, whose outputs are the predictions of sentiments
353   towards targeted aspects of entities.
354   The construction of inputs can be applied to the TABSA and MEABSA directly. For the ABSA
355   task, there is no entity mentioned in the post, the underlined part in the constructed question
356   template, which is "What is the sentiment towards the [aspect] **of [entity]**?", will be omitted.
357   This makes the BERT-based basic model be able to deal with all the ABSA, TABSA and
358   MEABSA tasks.
359   **Fusion Strategy**
360   Ensemble methods can improve the predictive performance of a single model by training
361   multiple models and combining their predictions. The weighting method is one of the effective
362   strategies to fuse outputs, which assign weights to each basic model to combine the final decision
363   (Sagi & Rokach, 2018), including simple averaging and weighted averaging (Zhou, 2021). We
364   follow the strategy of simple averaging and combine the data augmented CEA with BERT to be
365   the final model. We train the two models separately, and ensemble their predictions by taking the
366   sentiment polarity with the largest averaged predicted probability as the final output. For a given
367   post $Post_m$, the fine-grained sentiment prediction towards $aspect_j$ of $entity_i$, denoted as
368   $y_{entity_i}^{aspect_j}$, is calculated as Equation (5) shows.

369   $$P(c_i) = 0.5 \times P_{BERT}^{c_i}(Post_m, entity_i, aspect_j) + 0.5 \times P_{CEA}^{c_i}(Post_m, entity_i, aspect_j)$$

370   $$y_{entity_i}^{aspect_j} = argmax_i \, P(c_i) \qquad\qquad (5)$$

371   where $c_i \in \{positive, neutral, negative\}$, $P(c_i)$ represents the probability that the sentiment is
372   $c_i$, $P_{model}^{c_i}(Post_m, entity_i, aspect_j)$ represents the predicted probability of the sentiment $c_i$
373   towards  $aspect_j$ of $entity_i$ in $Post_m$ by the basic model BERT or data augmented CEA.
374   **Time Complexity Analysis**
375   Compared with existing deep learning-based models, our proposed PEA model involves entity
376   replacement, dual noise injection and prediction fusion as additional modules. The analysis of
377   time complexity for these three parts is described as follows.
378   For entity replacement, we calculated the selected probability for every entity, whose time
379   complexity is $O(E)$, where $E$ is the total number of entities in the dataset. We then traversed
380   every instance and conduct entity replacement, whose time complexity is $O(N)$, where $N$ is the
381   number of instances in the data set. The total time complexity of entity replacement is $O(E)$
382   $+ O(N)$.
383   For dual noise injection, we traversed every token in each instance to find the tokens referring to
384   entity and aspect, whose time complexity is $O(T)$, where $T$ is the length of each instance. We

385  added dual noises on all instances, whose time complexity is also $O(N)$. The total time
386  complexity of dual noise injection is $O(T) \times O(N)$.
387  For prediction fusion, we fused the prediction with the weighted summation operation on every
388  category for each instance, whose time complexity is $O(c) \times O(N)$, where $c$ is the number of
389  categories of sentiments.
390  The total time complexity of extra operations in our proposed PEA model is $(O(E) + O(N))$
391   $+ (O(T) \times O(N)) + (O(c) \times O(N))$.
392

## Experiments and Analysis

394  In this section, we introduce the experimental settings and results to validate the effectiveness of
395  our PEA model.

**Experimental Settings**

397  We evaluate four benchmark datasets of three tasks, including datasets in two languages: English
398  and Chinese. Statistics of the used datasets are displayed in Table 3.

399  • Restaurant and Laptop are two datasets from SemEval 2014 (Pontiki et al., 2014) for ABSA.
400  Both datasets are reviews in English and each review contains aspects and corresponding
401  sentiment polarities, including positive, negative and neutral.

402  • SentiHood is a widely used dataset for TABSA (Saeidi et al., 2016). It consists of 5,215
403  sentences in English, and 3,862 of which contain a single aspect, the rest contains multiple
404  aspects. Each sentence is annotated with a list of tuples, which are aspect, given entity and
405  corresponding sentiment polarity, including positive and negative. The whole dataset is split into
406  train, validation and test set.

407  • BabyCare is a large public dataset for MEABSA (Yang et al. 2018). It consists of babycare
408  reviews in Chinese and each review is in the format of a list of tuples, which are context, aspects,
409  corresponding entities and sentiment polarities, including positive, negative and neutral. The
410  whole dataset is split into train, validation and test set.

411  **Common settings.** For the BERT and CEA models, we use default parameters. For all English
412  datasets, we use BERT-Base English models[1] and 6B300d GloVe (Pennington, Socher &
413  Manning, 2014) word embeddings[2]. For the Chinese dataset, we use BERT-Base Chinese and the
414  same word vectors provided by (Yang et al., 2018). For multi-word entity terms and aspect
415  terms, we follow the preprocessing in previous works (Yang et al., 2018; Song et al., 2019; Yang
416  et al., 2019). We use the average vectors of all the words in the entity/aspect term as the
417  entity/aspect term vectors.

418  **Task specific settings.** For ABSA task, the *Restaurant* and *Laptop* datasets are used for
419  experiments. Because there is no entity in these datasets, so entity replacement in data
420  augmentation is removed when implementing PEA. For TABSA task, the *SentiHood* dataset is
421  used for experiments. Because aspect location is not given in this dataset, aspect noise injection

---

[1] https://github.com/google-research/bert.
[2] https://nlp.stanford.edu/projects/glove/

422 is removed in this task. For MEABSA task, the *BabyCare* dataset is used for experiments. When
423 implementing PEA, both entity replacement and noise injection are remained in this task.
424 **Data augmentation settings.** We perform entity replacement on the training data for the whole
425 dataset and merge the pseudo instances with original instances. According to the proposed entity
426 replacement method, those entities, which are low-resource in the original training set, have a
427 higher probability to be chosen for replacement. Table 4 lists the top 10 low-resource entities in
428 the *BabyCare* dataset, and displays the number of instances that belong to every category for
429 both the original training set and the entity-replaced dataset. We can observe that, for those low-
430 resource entities, such as "Kabrita", the number of negative and neutral instances has
431 significantly increased by using entity replacement. This can help relieve both the low-resource
432 and polarity bias problems.
433 For noise injection, $\mu$ and $\sigma$ are two parameters to be determined. We follow the common setting
434 in previous works (Kim, Roh & Kim, 2019) for $\mu$, which is $\mu = 0$. For $\sigma$, we conduct
435 experiments on all four datasets with $\sigma$ ranging from 0.01 to 0.4 to quantify the noise level.
436 Experimental results are in Fig. 4.
437 The $x$-axis refers to different values of $\sigma$, the $y$-axis refers to the Macro-F1 performance. Four
438 lines with different kinds of marks refer to the results of four datasets. Experimental results show
439 that when $\mu = 0$ and $\sigma = 0.05$, noise injection achieves the utmost performance on all tasks. We
440 use this setting in the following experiments.
441 **Model implementation settings.** We implement our proposed model with TensorFlow 2.1,
442 Python 3.7. The device we used consists of CPU (E5 2630 v4), GPU (1080ti * 4) and RAM
443 (256G). We compare our model with the state-of-the-art baselines on 3 tasks predicting
444 sentiment towards entities and aspects.
445 **Results**
446 *Accuracy* and *Marco-F1* score are two main-stream metrics in most sentiment analysis research,
447 where *Marco-F1* is the *F1* score averaged over all the classes. In the following experiments,
448 *Marco-Precision, Macro-Recall and AUC score are also used according to different tasks.*
449 **Results on the ABSA Task.** We evaluate the English benchmark datasets[3] *Restaurant* and
450 *Laptop* for the ABSA task. We compare with the published state-of-the-art baselines, including
451 Target-Dependent Long Short-Term Memory (TD-LSTM) (Tang et al., 2016), MemNet (Tang,
452 Qin & Liu, 2016), Attention-based LSTM with Aspect Embedding (ATAE-LSTM) (Wang et al.,
453 2016), Interactive Attention Network (IAN) (Ma et al., 2017), Recurrent Attention on Memory
454 (RAM) (Chen et al., 2017), Transfer Capsule Network (TransCap) (Chen & Qian, 2019), Aspect-
455 specific Graph Convolutional Network (ASGCN) (Zhang, Li & Song, 2019), and Capsule
456 Network with Interactive Attention (IACapsNet) (Du et al., 2019). Following the former
457 research, *Accuracy* and *Marco-F1* are evaluated for both datasets, *Marco-Precision and Macro-*
458 *Recall are also reported*. There is no entity in the dataset, so entity replacement in data
459 augmentation is removed. Results on two ABSA datasets are shown in Table 5.
460 We can have the following observations:

---

[3] http://alt.qcri.org/semeval2014/task4/

461  (1) by observing the accuracy and F1 performance, two Capsule Network-based models
462  TransCap and IACapsNet are much better than other previous baselines. This is because the key
463  components of TransCap and IACapsNet are recurrent neural works and attention mechanisms.
464  It shows that the RNN-based model has advantages in predicting fine-grained sentiments over
465  conventional methods.
466  (2) by observing the precision and recall on both datasets, the recall scores of most models
467  include TD-LSTM, ATAE-LSTM, IAN, RAM and ASGCN are much worse, while PEA can
468  have better performance.
469  (3) compared with all the baselines, our proposed model PEA achieves significant improvements
470  on both datasets. The experimental results show the PEA model is superior to other baselines in
471  the ABSA task under all evaluation metrics.
472  **Results on the TABSA Task.** We evaluate the English benchmark dataset *SentiHood* for the
473  TABSA task. It consists of 5,215 sentences, 3,862 of them contain a single target, and the
474  remainder multiple targets. We compare with all the published state-of-the-art baselines,
475  including Logistic Regression (LR) (Saeidi et al., 2016), LSTM+TA+SA (Ma, Peng & Cambria,
476  2018), SenticLSTM (Ma, Peng & Cambria, 2018), Dmu-Entnet (Liu, Cohn & Baldwin, 2018),
477  RE+Delayed-memory (Liang et al., 2019), BERT-pair-QA-B and BERT-pair-QA-M (Sun,
478  Huang & Qiu, 2019). Following the former research in the TABSA task, *Accuracy* and *AUC* are
479  usually reported and used as evaluation metrics, in the paper, *Marco-Precision, Macro-Recall*
480  *and Marco-F1 are also reported*. Results on TABSA are presented in Table 6.
481  We can have the following observations:
482  (1) BERT-pair-QA-M and BERT-pair-QA-B are the previous state-of-the-art models. Compared
483  with other none-BERT based baselines, BERT-pair-QA-M and BERT-pair-QA-B outperform the
484  LR, LSTM+TA+SA, SenticLSTM, Dmu-Entnet and RE+Delayed-memory models in both
485  accuracy and AUC score. This result shows the effectiveness of the pre-trained language model
486  for fine-grained sentiment analysis.
487  (2) compared with two BERT-based baselines, our proposed PEA achieves further improvement
488  in most evaluation metrics. This may be because the prediction of PEA comes from both data
489  augmented CEA and BERT, which helps ensemble the predictions of two basic models.
490  (3) different from the performance in ABSA and MEABSA, the improvement of PEA in the
491  TABSA task seems slightly in accuracy and AUC score, this may be because aspect location is
492  not given in this dataset (but given in other tasks), therefore, aspect noise injection is removed
493  for this experiment. So we have conducted a statistical analysis test in the following section to
494  show the performance difference between the two models is statistically significant.
495  **Results on the MEABSA Task.** We evaluate the Chinese benchmark dataset *BabyCare* for the
496  MEABSA task. We compare with all the published state-of-the-art baselines, including CEA
497  (Yang et al., 2018), DT-CEA (Yang et al., 2019), Cold-start Aware Deep Memory Network
498  (CADMN) (Song et al., 2019). These methods are exactly designed for this task. We also
499  compare with MemNet (Tang, Qin & Liu, 2016), ATAE-LSTM (Wang et al., 2016), IAN (Ma et
500  al., 2017), and their modified versions MemNet+, ATAE-LSTM+ and IAN+, which are used as

501   baselines in a recent MEABSA work (Song et al., 2019). We follow the designs introduced in
502   (Song et al., 2019): these three modified plus versions remain the basic model structure of
503   MemNet, ATAE-LSTM and IAN respectively. The additional entities in the MEABSA task are
504   treated as the aspects, and are added to the models in the same manner of aspects. These methods
505   are originally designed for the ABSA task, and they are often regarded as baselines in former
506   MEABSA research. Following the former research, *Accuracy* and *Marco-F1* are evaluation
507   metrics for this dataset, *Marco-Precision and Macro-Recall are also reported*. Table 7 displays
508   the comparisons between our model and baselines.
509   We can have the following observations:
510   (1) MemNet, ATAE-LSTM, and IAN in the first three lines only model aspects while ignoring
511   entity modeling. Their performances are worse than the plus versions MemNet+, ATAE-
512   LSTM+, and IAN+, which model the entity in the same manner as aspect, illustrating the
513   effectiveness of entity modeling in the MEABSA task.
514   (2) The CEA model combines the advantages of both attention-based LSTM and deep memory
515   networks, the former is the key component of ATAE-LSTM+ and the latter is the key component
516   of MemNet+. The performance of CEA is much better than ATAE-LSTM+ and MemNet+,
517   which reaches about 15% in accuracy. This shows that the CEA model has advantages in the
518   MEABSA task, and is more suitable to be chosen as an RNN-based basic model for PEA.
519   (3) DT-CEA and CADMN are two extension models based on CEA. DT-CEA incorporated
520   dependency information to improve CEA. CADMN used a frequency-guided attention
521   mechanism to improve CEA. The performance of CADMN and DT-CEA are comparable to each
522   other and are little better than CEA.
523   (4) compared with all the baselines, our proposed method PEA achieves significant improvement
524   under all evaluation metrics. Compared with the previous state-of-the-art CADMN model, the
525   improvements of PEA reach about 4% in accuracy and 5% in F1. The MEABSA is the most
526   challenging fine-grained sentiment analysis task, this experimental result shows PEA has a
527   significant advantage in the MEABSA task.
528   **Statistical Analysis Test.** Refer to the previous works (Li et al., 2020), we conduct McNemars
529   test as the statistical analysis test to further show the statistical difference between two models. $p$
530   -value is the significance level, which means the performance difference between the two models.
531   If the estimated $p$-value is lower than 0.05, the performance difference between the two models
532   is statistically significant. Table 8 displays the $p$-values between PEA and other models on three
533   sentiment analysis tasks respectively.
534   We can observe that the performance differences between PEA and other baselines are
535   statistically significant in all tasks, which show the effectiveness of the proposed PEA model
536   from the perspective of statistical analysis. For example, in the TABSA task, the improvement of
537   PEA compared with BERT-pair-NLI-M is not very high in accuracy, which is 94.3% vs 93.8%
538   in Table 6. In the statistical analysis test, the estimated $p$-value between PEA and BERT-pair-
539   NLI-M is 0.0174. According to the definition of $p$-value, it shows that the performance
540   difference between BERT-pair-NLI-M and PEA is statistically significant. Additionally, by

541 observing Table 7 and Table 8 together, we can find PEA has significant advantages in the most
542 challenging MEABSA task.
543

544 **Ablation Study**
545 Experimental results so far show that the PEA approach is superior to the baselines on all the
546 ABSA, TABSA and MEABSA on selected datasets. Because PEA consists of data augmented
547 CEA and BERT, we would like to further investigate the effectiveness of each part in the model.
548 A case study is also introduced in this section.
549 **Effectiveness of Components in PEA**
550 Ablation study is used to show how each part of the model affects the performance by removing
551 them. We conduct experiments on all four datasets of three tasks for comparisons. Experimental
552 results are as Table 9 shows.
553 The proposed PEA model integrates data augmented CEA and BERT. Because entity
554 replacement and noise injection are applied to data augmented CEA, we use CEA,
555 CEA+EntityReplacement (CEA+ER for short) and CEA+EntityReplacement+NoiseInjection
556 (CEA+ER+NI for short) respectively for ablation study to show the effectiveness of applying
557 two data augmentation techniques. The BERT-based model is also used for comparisons in
558 ablation studies.
559 We can have the following observations from Table 9:
560 (1) comparing CEA and CEA+ER, we can find involving entity replacement can have
561 improvement on MEABSA and TABSA tasks. We also counted the number of instances for
562 every entity based on the original training set and the entity-replaced dataset. The statistics are
563 demonstrated with the box plot in Fig. 5.
564 It shows that using the proposed entity-replacement method can significantly increase the
565 number of instances for low-resource entities, and all entities have at least 252 instances for
566 training. For ABSA, there is no entity provided in the dataset, so the entity replacement
567 procedure is removed.
568 (2) by adding noise injection, the CEA+ER+NI model achieves about 1.3% improvement on the
569 Restaurant dataset over the CEA+ER model, and achieves slight improvement on other datasets.
570 These observations show that using entity replacement and noise injection can bring positive
571 impacts on fine-grained sentiment analysis. This may be because using data augmentation can
572 increase the number of training instances, especially for low-resource entities and aspects, and
573 help overcome polarity bias.
574 (3) by comparing the performance of PEA with the BERT-based model and data augmented
575 CEA model, PEA achieves the best performance in most cases. The strength of BERT-based
576 model is that it makes use of a huge amount of unlabeled data by pre-training, but it also has
577 weaknesses. The BERT model depends on the Transformer (Vaswani et al., 2017), which further
578 mainly relies on its self-attention mechanism. It has been suggested that self-attention has
579 limitations that it cannot process input sequentially (Dehghani et al., 2018; Hao et al., 2019; Shen
580 et al., 2018; Hahn, 2020). Such a weakness is just the strength of recurrent neural networks,

581   which is one of the core components in CEA. Our model PEA combines the advantages of both
582   and performs the best in most cases. To better understand the strengths and weaknesses of data
583   augmented CEA and BERT, we carry out a case study in the next section.
584   **Case Study**
585   We give empirical validation on the strengths and weaknesses of two basic models, including the
586   BERT-based model and data augmented CEA, by a further case study on misclassifications of
587   both models. We test on the most challenging task MEABSA and use the corresponding
588   *Babycare* dataset for the case study. To show the stability of the models rather than the
589   occasionality, we have trained the BERT-based model, the Data augmented CEA model and the
590   PEA model five times. The predictions of two representative examples are as Table 10 shows.
591   For example 1, the BERT-based model makes the same misclassification on the inputs five times
592   and the data augmented CEA model achieves the correct predictions. Example 2 is just the
593   opposite. Such stable misclassifications reveal the defects of both models.
594   The first example has a special pattern: the coreference structure of "...the former..., ...the
595   latter...". The second example consists of two simple sentences. Correctly predicting the first
596   example need the ability of global sequence or structure understanding which is the advantage of
597   recurrent neural networks. The recurrent neural network is one of the core components of CEA.
598   Correctly predicting the second example need the ability of local attention which is the
599   advantage of self-attention, which is the core component of the BERT-based model. PEA fuses
600   the prediction with both BERT-based model and data augmented CEA model based on ensemble
601   methods, which make the correct prediction on both examples. This case study further helps
602   illustrate the value and necessity of ensembling two basic models.
603   We also give the third example in Table 10, where all the CEA, BERT-based model and PEA
604   made the wrong prediction. The gold output should be negative, but all models predicted it as
605   neutral. The possible reason is that there are no aspect terms directly towards the target entity
606   'Kao', which cause the model to give the prediction as neutral.
607

608   # Results on Challenging Conditions
609   There are two challenges in sentiment prediction towards entities and aspects: the low-resource
610   problem and the polarity bias problem. In this section, we evaluate the negative effect of
611   challenges and the ability of models to solve them.
612   **Results on Extreme Low-Resource Conditions**
613   To further test the model's performance under extreme low-resource conditions, we randomly
614   selected 5%, 10%, 20%, and 50%, each time, from the original dataset as our training dataset. All
615   tests are performed under the most challenging *Babycare* dataset. Experimental results are as
616   Fig. 6 shows.
617   The *x*-axis refers to the percentage of data used for training, the *y*-axis refers to the Macro-F1 of
618   different models. ER and NI are the abbreviations of entity replacement and noise injection. We
619   can have the following observations from Fig. 6.

620   (1) for all models, as the percentage of the training set used decreases, the models' performance
621   drops significantly, which further illustrates the significance of the low-resource problem on
622   sentiment prediction.
623   (2) CEA+ER outperforms the CEA model under all the low-resource conditions, which shows
624   the effectiveness of using entity replacement. By using noise injection, the CEA+ER+NI
625   achieves further improvements over CEA and CEA+ER.
626   (3) for the BERT-based model, when the resource is extremely low, the BERT-based model
627   deteriorates sharply. For example, when 5% of data is used for training, the Macro-F1 of BERT-
628   based model and PEA is 57.16% vs 64.37%. This shows that the combination of the data
629   augmented CEA and BERT-based model for PEA can boost the stability of the model. (4) the
630   dotted line in red refers to the baseline results with 100% data for training, we can observe from
631   Fig. 6 that when only 20% data are used for training, the proposed PEA can achieve a similar
632   performance of the CEA model with full-resource data for training. With the size of training data
633   becoming larger, the improvement of PEA becomes more obvious. This shows the PEA model,
634   which combines data augmented CEA with BERT-based model, has advantages under low-
635   resource conditions.
636   **Results on Evident Polarity Biased Conditions**
637   Polarity bias occurs when sentiment polarity distribution towards an entity is not uniform.
638   Polarity bias reduces the performance when sentiments towards an entity diverge in the training
639   set and in the test set (e.g. 70% of sentiment towards entity A are positive in the training set
640   while 60% of which are negative in the test set). We create a new test set named EPB test set,
641   which consists of all the instances with entities polarity biased from the original test set. Using
642   the *BabyCare* test set, we find entities in 30% of data (1070 out of 3677) have the evident
643   polarity bias problem. Experimental results are as Table 11 shows, the last column displays the
644   decline between the performance on the Original test set and EPB test set.
645   After comparing the sentiment prediction results from using the evident polarity biased data with
646   the results from using the origin data, we have the following observations:
647   (1) the performance of all models has varying degrees of decline on the polarity biased EPB
648   dataset. This shows the polarity bias problem is one of the challenges in fine-grained sentiment
649   analysis.
650   (2) comparing CEA and CEA+DA, the performance on the EPB test dataset is close to the
651   performance on the original test set. This is because data augmentations can relieve the polarity
652   bias problem by providing plenty, omni-polar sentiment training data, and reduced the variance
653   of test results to offer more stable performance. This shows applying data augmentations can
654   address the polarity bias problem in fine-grained sentiment analysis and make the model more
655   generality.
656   (3) comparing CEA and the BERT-based model, the performance on the original test set of the
657   BERT-based model has a significant improvement than that of CEA. (4) PEA achieves the best
658   performance on the original test set, and relieves the polarity problem on the EPB test at the

659  same time, which also shows the necessity and effectiveness of using the ensemble methods to
660  fuse the predictions of CEA and BERT based models with data augmentations.
661

## Conclusions

663  In this paper, we developed the PEA model, which unified the ABSA, TABSA, and MEABSA
664  tasks together for the first time and provided an all-in-one solution to interpret consumers'
665  opinions on all kinds of social media platforms. For the first time, we analysed the effect of the
666  sentiment polarity bias problem in these tasks. Most importantly, we created two innovative,
667  task-specific methods to alleviate the low-resource problem and the polarity bias problem, not
668  only getting promising experimental results, but also providing inspiration for successors to
669  make more contributions in this area. For future work, there are two possible extensions worth
670  considering. The first one is to look for new ways to combine pre-trained language models with
671  RNN-based models, to integrate both advantages. The second one is to further investigate more
672  types of fine-grained sentiment analysis, and propose unified models handling various fine-
673  grained sentiment-related tasks, for example, emotion cause analysis.
674

## Acknowledgements

676  We would like to sincerely thank our       colleagues, Professor Xiao-jun Wu, and Jun Sun
679  from Jiangnan University, Professor Jun-     yuan Xie, and Chong-jun Wang from Nanjing
680  University, for their kind support and       guidance.

682

## References

684  **Acheampong F, Nunoo-Mensah H, Chen W**. Transformer models for text-based emotion
685  detection: a review of BERT-based approaches. Artificial Intelligence Review, 2021: 1-41.
686  **Chen P, Sun Z, Bing L, Yang W**. Recurrent attention network on memory for aspect sentiment
687  analysis. in Proceedings of EMNLP, Copenhagen, Denmark, 2017: 452-461.
688  **Chen Z, Qian T**. Transfer capsule network for aspect level sentiment classification. in
689  Proceedings of ACL, Florence, Italy, 2019: 547-556.
690  **Chen Z, Qian T**. Relation-aware collaborative learning for unified aspect-based sentiment
691  analysis. in Proceedings of ACL, Seattle, Washington, USA, 2020: 3685-3694.
692  **Cheng Y, Tu Z, Meng F, Zhai J, Liu Y**. Towards Robust Neural Machine Translation. in
693  Proceedings of ACL, Melbourne, Australia, 2018: 1756-1766.
694  **Dai J, Yan H, Sun T, Liu P, Qiu X**. Does syntax matter? A strong baseline for Aspect-based
695  Sentiment Analysis with RoBERTa. in Proceedings of NAACL-HLT, online, 2021: 1816-1829.
696  **Dai X, Adel H**. An Analysis of Simple Data Augmentation for Named Entity Recognition. in
697  Proceedings of COLING, Barcelona, Spain, 2020: 3861-3867.
698  **Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser Ł**. Universal Transformers. in
699  Proceedings of ICLR, Vancouver, BC, Canada. 2018.

700 **Devlin J, Chang M.W, Lee K, Toutanova K**. Bert: Pre-training of deep bidirectional
701 transformers for language understanding. in Proceedings of NAACL-HTL, New Orleans,
702 Louisiana, USA, 2018: 4171–4186.
703 **Du C, Sun H, Wang J, Qi Q, Liao J, Xu T, Liu M**. Capsule network with interactive attention
704 for aspect-level sentiment classification. in Proceedings of EMNLP-IJCNLP, Hong Kong, China,
705 2019: 5492-5501.
706 **Fadaee M, Bisazza A, Monz C**. Data Augmentation for Low-Resource Neural Machine
707 Translation. in Proceedings of ACL, Vancouver, Canada, 2017: 567-573.
708 **Gupta R**. Data augmentation for low resource sentiment analysis using generative adversarial
709 networks. in Proceedings of ICASSP, Brighton, UK, 2019: 7380-7384.
710 **Hahn M**. Theoretical limitations of self-attention in neural sequence models. Transactions of the
711 Association for Computational Linguistics, 2020, 8: 156-171.
712 **Han H, Zhang Y, Zhang J, Yang J, Zou X**. Improving the performance of lexicon-based
713 review sentiment analysis method by reducing additional introduced sentiment bias. PloS one,
714 2018, 13(8): e0202523.
715 **Hao J, Wang X, Yang B, Wang L, Zhang J**. Modeling Recurrence for Transformer. in
716 Proceedings of NAACL-HTL, Minneapolis, MN, USA, 2019: 1198-1207.
717 **Hong H, Song J**. Utilizing BERT for Detecting Aspect Categories on TABSA via Adjusting
718 Self-attention among Words. in Proceedings of ICHCI, Sanya, China, 2020: 66-70.
719 **Hou Y, Liu Y, Che W, Liu T**. Sequence-to-Sequence Data Augmentation for Dialogue
720 Language Understanding. in Proceedings of COLING, Santa Fe, New Mexico, USA, 2018:
721 1234-1245.
722 **Howard J, Ruder S**. Universal Language Model Fine-tuning for Text Classification. in
723 Proceedings of ACL, Melbourne, Australia, 2018: 328-339.
724 **Iqbal M, Karim A, Kamiran F**. Bias-aware lexicon-based sentiment analysis. in Proceedings of
725 SAC, Salamanca, Spain, 2015: 845-850.
726 **Khine W, Aung T**. Applying deep learning approach to targeted aspect-based sentiment analysis
727 for restaurant domain. in Proceedings of ICAIT, Jinan, China, 2019: 206-211.
728 **Kim H.Y, Roh Y.H, Kim Y.G**. Data augmentation by data noising for open-vocabulary slots in
729 spoken language understanding. in Proceedings of NAACL-HTL, Minneapolis, MN, USA, 2019:
730 97-102.
731 **Kiritchenko S, Mohammad S**. Examining Gender and Race Bias in Two Hundred Sentiment
732 Analysis Systems. in Proceedings of SIGLEX-SIGSEM, New Orleans, Louisiana, USA, 2018:
733 43-53.
734 **Li Z, Sun J, Wu X, Yin H.** Multiplication fusion of sparse and collaborative-competitive
735 representation for image classification. International Journal of Machine Learning and
736 Cybernetics, 2020, 11(10): 2357-2369.
737 **Liang B, Du J, Xu R, Li B, Huang H**. Context-aware Embedding for Targeted Aspect-based
738 Sentiment Analysis. in Proceedings of ACL, Florence, Italy, 2019: 4678-4683.

739   **Liu F, Cohn T, Baldwin T**. Recurrent Entity Networks with Delayed Memory Update for
740   Targeted Aspect-Based Sentiment Analysis. in Proceedings of NAACL, New Orleans,
741   Louisiana, USA. 2018: 278-283.
742   **Lu K, Mardziel P, Wu F, Amancharla P, Datta A**. Gender bias in neural natural language
743   processing. Logic, Language, and Security, 2020: 189-202.
744   **Ma D, Li S, Zhang X, Wang H**. Interactive attention networks for aspect-level sentiment
745   classification. in Proceedings of IJCAI, Melbourne, VIC, Australia, 2017: 4068-4074.
746   **Ma Y, Peng H, Cambria E**. Targeted aspect-based sentiment analysis via embedding
747   commonsense knowledge into an attentive LSTM. in Proceedings of AAAI, New Orleans, LA,
748   USA, 2018: 5876-5883.
749   **Peters M.E, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L**. Deep
750   Contextualized Word Representations. in Proceedings of ACL, Melbourne, Australia, 2018:
751   2227-2237.
752   **Pennington J, Socher R, Manning C. D**. Glove: Global vectors for word representation. in
753   Proceedings of EMNLP, Doha, Qatar, 2014: 1532-1543.
754   **Pontiki M, Papageorgiou H, Galanis D, Androutsopoulos I, Pavlopoulos J, Manandhar S**.
755   SemEval-2014 Task 4: Aspect Based Sentiment Analysis, in Proceedings of SemEval, Dublin,
756   Ireland, 2014: 27-35.
757   **Saeidi M, Bouchard G, Liakata M, Riedel S**. SentiHood: Targeted Aspect Based Sentiment
758   Analysis Dataset for Urban Neighbourhoods. in Proceedings of COLING, Osaka, Japan, 2016:
759   1546-1556.
760   **Sagi O, Rokach L**. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining
761   and Knowledge Discovery, 2018, 8(4): e1249.
762   **Song K, Gao W, Zhao L, Sun C, Liu X**. Cold-start aware deep memory networks for multi-
763   entity aspect-based sentiment analysis. in Proceedings of IJCAI, Macao, SAR, China, 2019:
764   5197-5203.
765   **Shen T, Zhou T, Long G, Jiang J, Pan S, Zhang C**. Disan: Directional self-attention network
766   for rnn/cnn-free language understanding. in Proceedings of AAAI, New Orleans, LA, USA,
767   2018: 5446-5455.
768   **Sun C, Huang L, Qiu X**. Utilizing BERT for Aspect-Based Sentiment Analysis via
769   Constructing Auxiliary Sentence. in Proceedings of ACL, Florence, Italy, 2019: 380-385.
770   **Tang D, Qin B, Feng X, Liu T**. Effective LSTMs for Target-Dependent Sentiment
771   Classification. in Proceedings of COLING, Osaka, Japan, 2016: 3298-3307.
772   **Tang D, Qin B, Liu T**. Aspect Level Sentiment Classification with Deep Memory Network. in
773   Proceedings of EMNLP, Austin, Texas, USA, 2016: 214-224.
774   **Thelwall M**. Gender bias in machine learning for sentiment analysis. Online Information
775   Review, 2018, 42(1): 45-57.
776   **Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Polosukhin I**.
777   Attention is all you need. in Proceedings of NeurIPS, Long Beach, CA, USA, 2017: 6000-6010.

778 **van Aken B, Winter B, Löser A, Gers F**. How does bert answer questions? a layer-wise
779 analysis of transformer representations. In Proceedings of CIKM, Beijing, China, 2019: 1823-
780 1832.
781 **Wan H, Yang Y, Du J, Liu Y, Qi K, Pan J**. Target-aspect-sentiment joint detection for aspect-
782 based sentiment analysis. in Proceedings of AAAI, New York, New York, USA, 2020: 9122-
783 9129.
784 **Wang K, Shen W, Yang Y, Quan X, Wang R**. Relational Graph Attention Network for Aspect-
785 based Sentiment Analysis. in Proceedings of ACL, Online Conference, 2020: 3229-3238.
786 **Wang W.Y, Yang D**. That's so annoying!!!: A lexical and frame-semantic embedding based
787 data augmentation approach to automatic categorization of annoying behaviors using# petpeeve
788 tweets. in Proceedings of EMNLP, Lisbon, Portugal, 2015: 2557-2563.
789 **Wang Y, Huang M, Zhu X, Zhao L**. Attention-based LSTM for aspect-level sentiment
790 classification. in Proceedings of EMNLP, Austin, Texas, USA, 2016: 606-615.
791 **Wu Z Ong D**. Context-guided bert for targeted aspect-based sentiment analysis. in Proceedings
792 of AAAI, New York, New York, USA, 2020: 1-9.**Xia M, Kong X, Anastasopoulos A, Neubig**
793 **G**. Generalized Data Augmentation for Low-Resource Translation. in Proceedings of ACL,
794 Florence, Italy, 2019: 5786-5796.
795 **Xu B, Wang X, Yang B, Kang Z**. Target embedding and position attention with lstm for aspect
796 based sentiment analysis. in Proceedings of ICMAI, Chengdu, China, 2020: 93-97.
797 **Yang J, Yang R, Wang C, Xie J**. Multi-entity aspect-based sentiment analysis with context,
798 entity and aspect memory. in Proceedings of AAAI, New Orleans, LA, USA, 2018: 6029-6036.
799 **Yang J, Yang R, Lu H, Wang C, Xie J**. Multi-entity aspect-based sentiment analysis with
800 context, entity, aspect memory and dependency information. ACM TALLIP, 2019, 18(4): 1-22.
801 **Ye Z, Li Z**. A variant of recurrent entity networks for targeted aspect-based sentiment analysis.
802 in Proceedings of ECAI, Santiago de Compostela, Spain, 2020: 2268-2274.
803 **Zhao J, Wang T, Yatskar M, Ordonez V, Chang K. W**. Gender Bias in Coreference
804 Resolution: Evaluation and Debiasing Methods. in Proceedings of NAACL, New Orleans,
805 Louisiana, USA, 2018: 15-20.
806 **Zhang C, Li Q, Song D**. Aspect-based Sentiment Classification with Aspect-specific Graph
807 Convolutional Networks. in Proceedings of EMNLP-IJCNLP, Hong Kong, China, 2019: 4560-
808 4570.
809 **Zhang M, Qian T**. Convolution over Hierarchical Syntactic and Lexical Graphs for Aspect
810 Level Sentiment Analysis. in Proceedings of EMNLP, online, 2020: 3540-3549.
811 **Zhou Z**. Ensemble learning. Machine Learning. Springer, Singapore, 2021: 181-210.
812 **Zhu P, Chen Z, Zheng H, Qian T**. Aspect aware learning for aspect category sentiment
813 analysis. ACM Transactions on Knowledge Discovery from Data. ACM TKDD, 2019, 13(6): 1-
814 21.

Table 1. The comparison between three tasks of sentiment prediction towards entities and aspects.

1    Table 1. The comparison between three tasks of sentiment prediction towards entities and aspects.

| | | ABSA | TABSA | MEABSA |
|---|---|---|---|---|
| **Definition** | **Given** | text and aspects mentioned | text, entity mentioned, and all kinds of aspects | text, entity mentioned, and aspect mentioned |
| | **Goal** | predict sentiment towards mentioned aspects | predict sentiment towards the combination of mentioned entity and all kinds of aspects | predict sentiment towards mentioned entity aspect combination |
| **Example1** | **Input** | \<**context**\> I've used MacBookPro, it's convenient. \</**context**\> \<**aspect** from ="27" to="37"\> convenience level\</**aspect**\> | \<**context**\> I've used MacBookPro, it's convenient. \</**context**\> \<**entity** from="10" to="20"\> MacBookPro \</**entity**\> \<**aspectlist**\> price, convenience level, battery, ... \</**aspectlist**\> | \<**context**\> I've used MacBookPro, it's convenient. \</**context**\> \<**entity** from="10" to="20"\>MacBookPro \</**entity**\> \<**aspect**\> from="27" to="37"\>convenience level\</**aspect**\> |
| | **Output** | (convenience level, positive) | (MacBookPro, price, none) (MacBookPro, convenience level, positive) (MacBookPro, battery, none) ... | (MacBookPro, convenience level, positive) |
| **Example2** | **Input** | \<**context**\>The battery of ThinkPad is very long. \</**context**\> \<**aspect** from ="4" to="11"\> battery\</**aspect**\> | \<**context**\> The battery of ThinkPad is very long. \</**context**\> \<**entity** from="15" to="23"\>ThinkPad \</**entity**\> \<**aspectlist**\> price, convenience level, battery, ... \</**aspectlist**\> | \<**context**\> The battery of ThinkPad is very long. \</**context**\> \<**entity** from="15" to="23"\>ThinkPad \</**entity**\> \<**aspect**\> from="4" to="11"\>battery\</**aspect**\> |
| | **Output** | (battry, positive) | (ThinkPad, price, none) (ThinkPad, convenience level, none) (ThinkPad, battery, positive) ... | (ThinkPad, battery, positive) |

Table 2. An example of entity replacement. The replacement maintains the same sentiment polarity and correct grammar.

1     Table 2. An example of entity replacement. The replacement maintains the same sentiment polarity and correct

2                                                        grammar.

| | | |
|---|---|---|
| **Original** | **post** | I've used MacBookPro, it's convenient. |
| | **entity** | MacBookPro |
| **New** | **post** | I've used Thinkpad, it's convenient. |
| | **target entity** | Thinkpad |

Table 3. Statistics of used datasets.

Table 3. Statistics of used datasets.

| Dataset | Language | Training set | Validation set | Test set | Task ABSA | Task TABSA | Task MEABSA |
|---|---|---|---|---|---|---|---|
| Restaurant | English | 3,608 | - | 1,120 | ☞ | | |
| Laptop | English | 2,328 | - | 638 | ☞ | | |
| SentiHood | English | 3,650 | 522 | 1,043 | | ☞ | |
| BabyCare | Chinese | 29,354 | 3,682 | 3,677 | | | ☞ |

Table 4: Definition of the confusion matrix.

1   Table 4. Top 10 low-resource entities in the BabyCare dataset, with the number of instances that belong to
2                  every polarity category for both the original training set and entity-replaced dataset.

|  | the original training set | | | the entity-replaced dataset | | |
|---|---|---|---|---|---|---|
| Entity | Negative | Neutral | Positive | Negative | Neutral | Positive |
| 佳贝艾特(Kabrita) | 0 | 0 | 26 | 69 | 243 | 361 |
| 可瑞康(Karicare) | 0 | 0 | 17 | 108 | 382 | 527 |
| 君乐宝(JunLeBao) | 0 | 0 | 15 | 121 | 409 | 571 |
| 咔哇熊(Cowala) | 0 | 0 | 14 | 144 | 446 | 642 |
| 多美滋(Dumex) | 1 | 0 | 64 | 22 | 84 | 219 |
| 太子乐(Happy Prince) | 1 | 0 | 19 | 102 | 306 | 485 |
| 奶粉(milk powder) | 0 | 0 | 19 | 102 | 304 | 479 |
| 欧贝嘉(OuBecca) | 1 | 0 | 19 | 86 | 305 | 459 |
| 百立乐(Natrapure) | 4 | 0 | 73 | 37 | 71 | 183 |
| 诺优能(Nutrilon) | 2 | 0 | 42 | 44 | 146 | 227 |

Table 5. Performance (%) on two datasets for the ABSA task, *Accuracy*, *Marco-Precision*, *Macro-Recall* and *Marco-F1*are reported.

1     Table 5. Performance (%) on two datasets for the ABSA task, *Accuracy*, *Marco-Precision*, *Macro-Recall* and

2                                    *Marco-F1* are reported.

| Models | Restaurant | | | | Laptop | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| TD-LSTM | 75.18 | 70.60 | 56.57 | 58.51 | 64.26 | 57.67 | 56.67 | 54.10 |
| MemNet | 77.32 | 69.87 | 64.38 | 64.61 | 68.65 | 63.58 | 63.62 | 62.69 |
| ATAE-LSTM | 74.38 | 67.43 | 57.28 | 58.32 | 66.14 | 61.22 | 58.97 | 56.91 |
| IAN | 76.16 | 67.43 | 59.31 | 60.56 | 65.20 | 61.64 | 58.54 | 54.08 |
| RAM | 76.07 | 72.07 | 58.65 | 59.59 | 68.03 | 64.03 | 63.86 | 60.82 |
| TransCap | 79.20 | 70.76 | 70.81 | 70.78 | 74.76 | 71.77 | 71.99 | 70.08 |
| ASGCN | 74.29 | 71.95 | 56.74 | 56.45 | 69.75 | 66.21 | 63.75 | 62.29 |
| IACapsNet | 81.79 | - | - | 73.40 | 76.80 | - | - | 73.29 |
| PEA(Our) | **84.82** | **80.41** | **76.31** | **78.14** | **78.68** | **74.43** | **76.60** | **75.07** |

Table 6. Performance (%) on the SentiHood dataset for the TABSA task, *Accuracy*, *Marco-Precision*, *Macro-Recall*, *Marco-F1* and *AUC* are reported.

1    Table 6. Performance (%) on the SentiHood dataset for the TABSA task, *Accuracy*, *Marco-Precision*, *Macro-*
2    *Recall*, *Marco-F1* and *AUC* are reported.

| Models | Accuracy | Precision | Recall | F1 | AUC |
|---|---|---|---|---|---|
| LR | 87.5 | - | - | - | 90.5 |
| LSTM+TA+SA | 86.8 | - | - | - | - |
| SenticLSTM | 89.3 | - | - | - | - |
| Dmu-Entnet | 90.2 | 74.8 | 76.3 | 75.5 | 94.8 |
| RE+Delayed-memory | 92.8 | - | - | - | 96.2 |
| BERT-pair-QA-B | 93.3 | - | - | - | 97.0 |
| BERT-pair-QA-M | 93.8 | 83.4 | **85.7** | 84.5 | 97.1 |
| PEA(Our) | **94.3** | **86.0** | 84.5 | **85.2** | **97.4** |

Table 7. Performance (%) on the BabyCare dataset for the MEABSA task, *Accuracy*, *Marco-Precision*, *Macro-Recall* and *Marco-F1* are reported.

1      Table 7. Performance (%) on the BabyCare dataset for the MEABSA task, *Accuracy*, *Marco-Precision*,

2      *Macro-Recall* and *Marco-F1* are reported.

| Models | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| MemNet | 62.74 | 59.81 | 48.84 | 46.13 |
| ATAE-LSTM | 66.09 | 58.47 | 49.68 | 47.75 |
| IAN | 61.93 | 41.71 | 47.04 | 43.73 |
| MemNet+ | 65.32 | 59.93 | 50.55 | 47.93 |
| ATAE-LSTM+ | 66.25 | 56.01 | 51.93 | 51.87 |
| IAN+ | 65.81 | 44.42 | 50.06 | 46.50 |
| CEA | 80.20 | 77.68 | 75.23 | 76.29 |
| DT-CEA | 81.74 | - | - | 78.23 |
| CADMN | 81.45 | - | - | 78.37 |
| PEA(Our) | **85.72** | **83.97** | **82.60** | **83.25** |

Table 8. $p$-value between PEA and other baselines on ABSA, TABSA and MEABSA tasks.

1        Table 8. *p*-value between PEA and other baselines on ABSA, TABSA and MEABSA tasks.

| ABSA Task | | |
|---|---|---|
| **Dataset** | **Restaurant** | **Laptop** |
| TD-LSTM | 1.4379e-14 | 1.4331e-14 |
| MemNet | 1.6116e-11 | 1.7323e-09 |
| ATAE-LSTM | 2.0494e-16 | 1.1331e-12 |
| IAN | 5.8819e-13 | 1.2102e-12 |
| RAM | 9.4895e-13 | 2.1595e-09 |
| TransCap | 1.1872e-06 | 0.0138 |
| ASGCN | 7.3462e-16 | 3.7338e-07 |
| **TABSA Task** | | |
| **Dataset** | **SentiHood** | |
| Dmu-Entnet | 6.7790e-41 | |
| BERT-pair-NLI-M | 0.0174 | |
| **MEABSA Task** | | |
| **Dataset** | **BabyCare** | |
| MemNet | 6.6475e-140 | |
| ATAE-LSTM | 8.3216e-113 | |
| IAN | 7.1802e-148 | |
| MemNet+ | 3.4485e-120 | |
| ATAE-LSTM+ | 1.1143e-114 | |
| IAN+ | 1.6552e-114 | |
| CEA | 2.2666e-20 | |

Table 9. Performance (%) of ablation study on 4 datasets.

<p style="text-align:center">Table 9. Performance (%) of ablation study on 4 datasets.</p>

| Ablation | BabyCare | | SentiHood | | Restaurant | | Laptop | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | F1 | Accuracy | AUC | Accuracy | F1 | Accuracy | F1 |
| CEA | 80.20 | 76.29 | 90.3 | 93.2 | 78.13 | 67.96 | 71.41 | 67.05 |
| CEA+ER | 80.78 | 77.33 | 91.1 | 93.3 | - | - | - | - |
| CEA+ER+NI | 81.06 | 77.55 | 91.3 | 94.0 | 79.45 | 70.31 | 71.83 | 67.22 |
| BERT-based model | 84.12 | 81.62 | 93.8 | 97.1 | 83.52 | 76.11 | 76.99 | 72.40 |
| PEA (our) | **85.72** | **83.97** | **94.3** | **97.4** | **84.82** | **78.14** | **78.68** | **75.07** |

Table 10. Case study on misclassifications of BERT-based and data augmented CEA model. The straight underlined words are entity terms and the wavy underlined words are aspect terms.

1    Table 10. Case study on misclassifications of BERT-based and data augmented CEA model. The straight
2         underlined words are entity terms and the wavy underlined words are aspect terms.

| | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| **Input** | I tried Pampers and Inherent. I dislike the smell of the former, and the latter leaks. | QiAn feels thin. I have brought BobDog, it will be delivered home. | I am too poor to afford Kao. My son and daughter are using XWW, it is cheap. |
| **Entity** | Pampers | BobDog | Kao |
| **Aspect** | Anti-leakage | Thickness | Cost |
| **BERT-based model output** | Negative | Neutral | Neutral |
| **Data augmented CEA output** | Neutral | Positive | Neutral |
| **PEA output** | Neutral | Neutral | Neutral |
| **Gold output** | Neutral | Neutral | Negative |

Table 11. *Macro-F1* and standard deviation of *Macro-F1* (in the brackets) on evident polarity biased (EPB) test set and original test set. "DA" is short for Data Augmentation.

1    Table 11. *Macro-F1* and standard deviation of *Macro-F1* (in the brackets) on evident polarity biased (EPB)
2                test set and original test set. "DA" is short for Data Augmentation.

| Models | EPB test set | Original test set | Decline on EPB |
|---|---|---|---|
| CEA | 0.7542(0.0123) | 0.7714(0.0040) | 1.72% |
| CEA+DA | 0.7753(0.0069) | 0.7768(0.0036) | 0.15% |
| BERT-based model | 0.8068(0.0070) | 0.8162(0.0040) | 0.94% |
| PEA | 0.8153(0.0069) | 0.8234(0.0061) | 0.81% |

# Figure 1

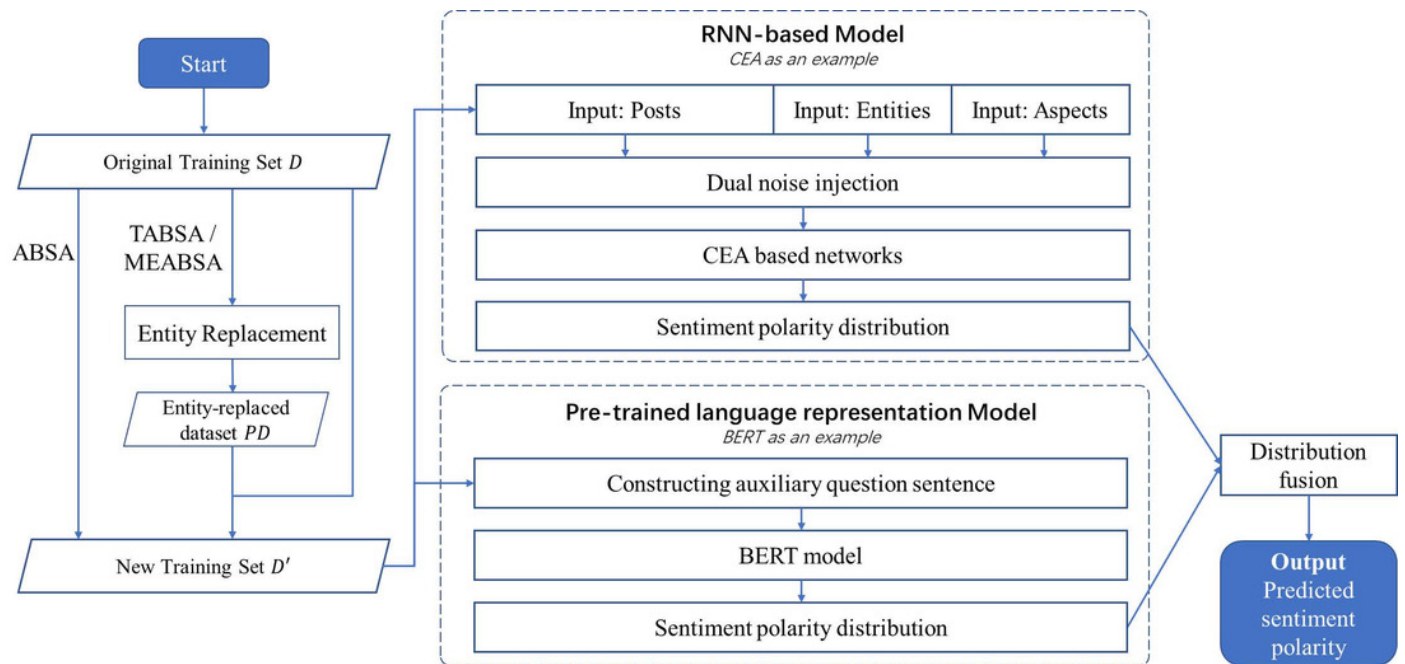
Figure 1. The graphical abstract of the PEA model.

# Figure 2

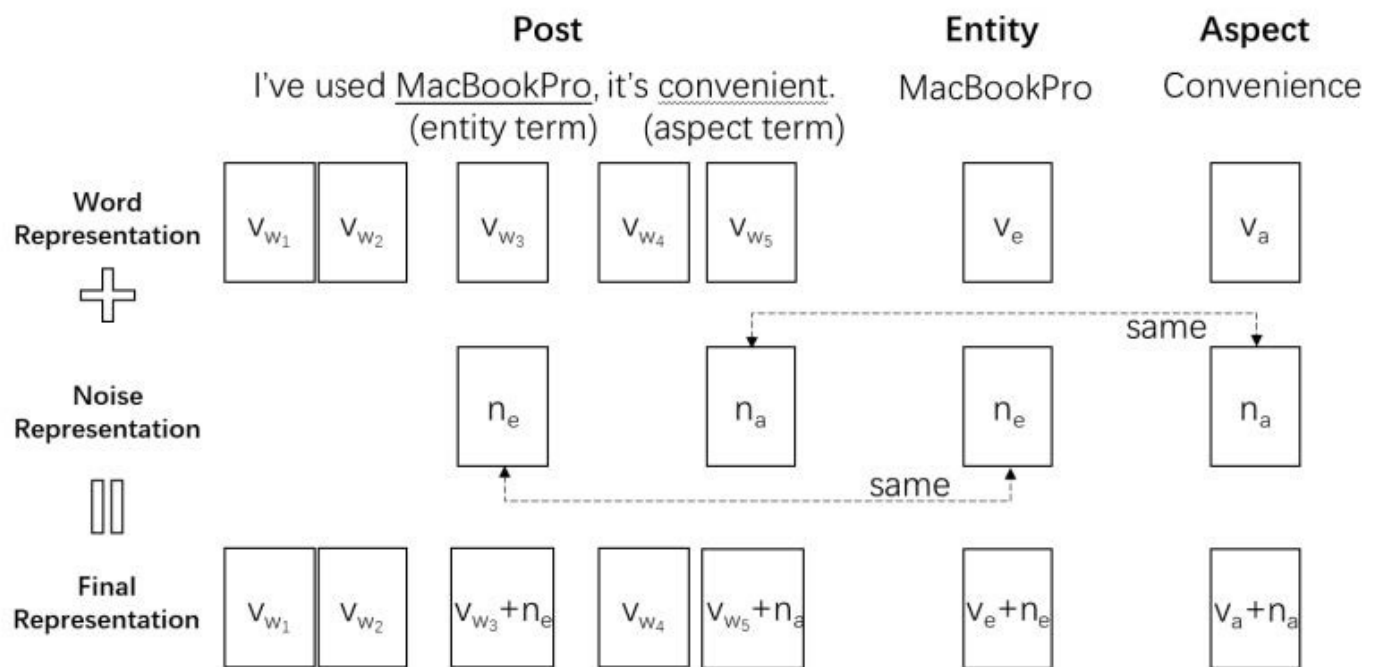
Figure 2. An example of dual noise injection.

# Figure 3

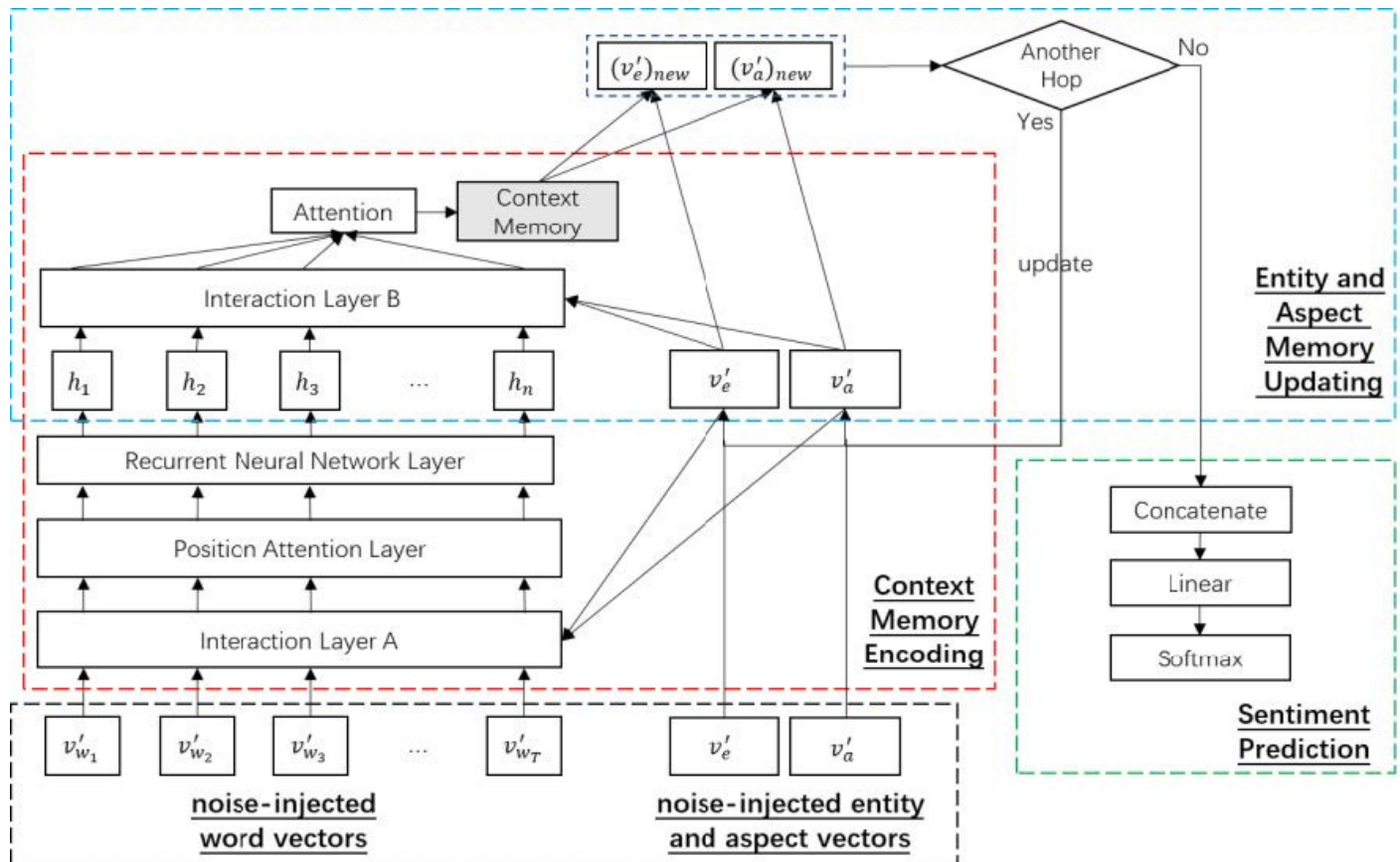
Figure 3. General structure of CEA with noise-injected vectors

# Figure 4

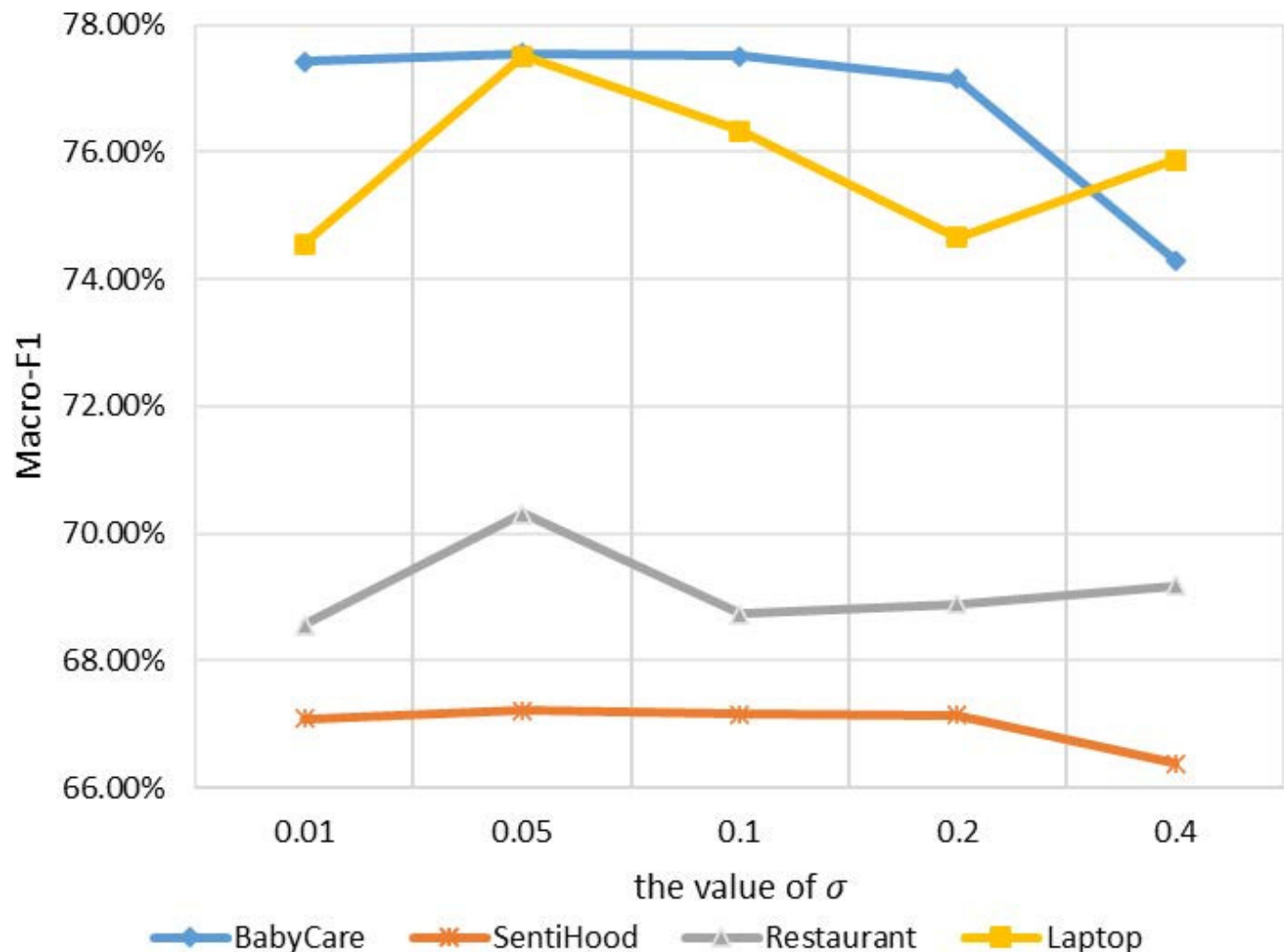
Figure 4. Macro-F1 performance on four datasets with different values of σ in noise infection

# Figure 5

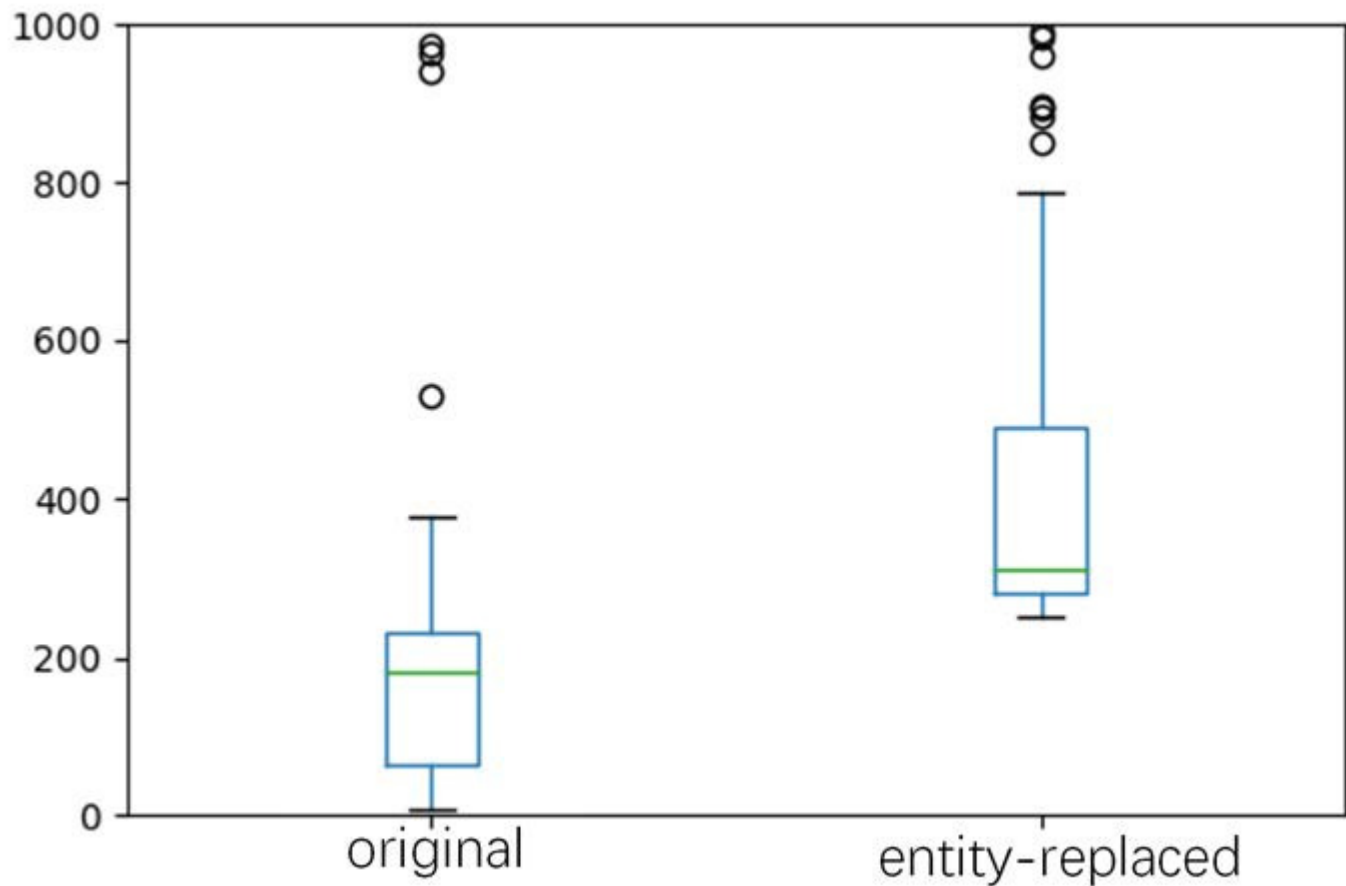
Figure 5. Box plot of the number of instances for every entity based on the original training and the entity-replaced dataset respectively.

# Figure 6

Figure 6. Performance on extreme low-resource conditions.