A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis

Anonymous ACL submission

Abstract

Sentiment analysis is an important task in natural language processing. In recent works, pre-trained language models are often used to achieve state-of-the-art results, especially when training data is scarce. It is common to fine-tune on the downstream task, usually by adding task-specific layers on top of the model. In this paper, we focus on aspect-based sentiment analysis, which involves extracting aspect term, category, and predicting their corresponding polarities. In particular, we are interested in few-shot settings. We propose to reformulate the extraction and prediction tasks into the sequence generation task, using a generative language model with unidirectional attention (GPT2 is used unless stated otherwise). This way, the model learns to accomplish the tasks via language generation without the need of training task-specific layers. Our evaluation results on the single-task polarity prediction show that our approach outperforms the previous state-of-the-art (based on BERT) on average performance by a large margins in few-shot and full-shot settings. More importantly, our generative approach significantly reduces the model variance caused by low-resource data. We further demonstrate that the proposed generative language model can handle joint and multi-tasking settings, unlike previous work. We observe that the proposed sequence generation method achieves further improved performances on polarity prediction when the model is trained via joint and multi-tasking settings. Further evaluation on similar sentiment analysis datasets, SST-2, SST-5 and OOS intent detection validates the superiority and noise robustness of generative language model in few-shot settings.

1 Introduction

Sentiment analysis (Pang et al., 2002; Turney, 2002; Chevalier and Mayzlin, 2006) aims at detecting the overall polarity of a user generated text, which describes the user opinion for an entity. However, user may express opinions about an entity at different granularity. For example, a user may give an overall rate about a restaurant service, and then explains fine-grained review about specific aspects, such as food quality, waiting time, waitress service, environment, etc. Aspect-based sentiment analysis task (Pontiki et al., 2014, 2016) aims at addressing this problem, where user sentiment is annotated at coarse and fine-grained levels. Moreover, user can express conflicting opinions for different aspects of an entity.

Traditionally, neural-based models are employed as a single-task model for aspect-based sentiment analysis (ABSA) task, similar to Machine Reading Comprehension task (MRC) (Rajpurkar et al., 2016). For example, a pre-trained BERT language model is fine-tuned for ABSA term polarity prediction (single-task) as a classifier. In this approach, a task-specific layer is fine-tuned for each downstream task, such as a layer for aspect term polarity classification, and a different layer for aspect term span extraction (Xu et al., 2019).

Recently, generative language models with unidirectional self-attention, which are pre-trained by causal language modeling loss (predicting next word given the history), have shown promising performance when fine-tuned on the downstream tasks (GPT2) (Radford et al., 2018). Using this approach, the language model learns the downstream task as language generation, where the task is represented as a serialized text. Moreover, Brown et al. (2020) proposed GPT3, a large-scale generative language model with few-shot ability. GPT3 learns to solve the downstream task by conditioning on few examples in the prompt, without any parameter update (in-context learning).

Motivated by the ability of the pre-trained generative language model (GPT2) for solving the downstream tasks in a generative manner, we propose a generative language model for ABSA task. The evaluation results indicate that the proposed approach achieves better performance with significantly lower variance compared to the previous state-of-the-art models (which are based on BERT pre-trained model) on few-shot and full-shot settings, for single-task polarity prediction of aspect term and aspect category. For example, using 1% (20 examples) of training data on restaurant domain for aspect term polarity prediction task, our proposed GPT2 model outperforms BERT-PT (Xu et al., 2019) by 9 points on average accuracy and reduced standard deviation by 6.2 points, as shown in Figure 1(a). Moreover, when fine-tuned on multiple tasks, such as aspect term extraction, term polarity, aspect category detection, and category polarity, the proposed model improved single-task performance, such as aspect term extraction (measured by F1 score).[1]

The contributions of our proposed generative language model are,

- A robust generative model on few-shot aspect-based sentiment analysis by reformulating the task as language generation. This allows us to use uni-directional language model with no additional head for the downstream tasks, which outperforms the previous state-of-the-arts on average performance by a large margin, with no additional pretraining on out-of-domain data (such as BERT-PT (Xu et al., 2019)).

- Our proposed generative model reduces variance in polarity prediction, caused by low resource data and random noise, in all few and full-shot settings by large value.

- Joint and multi-task training can further improve the single-task few-shot performances, such as aspect term extraction.

- More evaluation on similar sentiment analysis tasks (SST-2, SST-5, OOS intent detection) provides further evidence of the superiority and robustness of generative language model.

In the next sections, we discuss the proposed model and presents the evaluation results. In section 2, the previous state-of-the-arts are described. Section 3 explains the task of aspect-based sentiment analysis (ABSA) (section 3.1) followed by reformulating ABSA task as language generation (section 3.2). In section 4, the evaluation results for single, joint and multi-task settings are presented for SemEval14 (Pontiki et al., 2014) and SemEval16 (Pontiki et al., 2016) and SST-2, SST-5 and OOS intent detection datasets.

[1]Our code will be available at https://github.com/salesforce/absa_fewshot

2 Related Works

Sentiment analysis is characterized by three categorizes, i.e. document, sentence, and aspect level (Liu, 2012; Liu and Zhang, 2012; Cambria and Hussain, 2012). In this section, we review the previous models developed for aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004).

Earlier works on ABSA task focused on developing feature engineered models (Samha et al., 2014). Xu et al. (2018) proposed a model based on using convolutional neural network (CNN) for aspect term extraction task only. The approach uses two types of pre-trained embeddings, a general-purpose embedding and a domain-specific one. Then, a softmax classification layer is used to classify each word to identify aspect term start and end positions, or non-related words.

Li et al. (2019) proposed Multi-granularity Alignment Network (MGAN), a coarse-to-fine approach for single-task aspect term polarity prediction using recurrent neural network (RNN) (Hochreiter and Schmidhuber, 1997). They defined aspect category as coarse-level and aspect term as fine-level sentiments, and further leveraged high-resource out-of-domain data for pre-training. This way, the knowledge is transferred from coarse-grain domains (single-opinion prediction) to multi-grain domains (ABSA task).

With the advent of BERT (Devlin et al., 2018) as a pre-trained bidirectional language model, which presents a powerful contextualized word representation for the language understanding downstream tasks, several models are proposed for ABSA task using BERT as feature extraction. Xu et al. (2019) defined ABSA task as question answering (Rajpurkar et al., 2016), named Review Reading Comprehension (RRC), and used BERT as the base model, with separate heads for aspect term extraction (as span extraction) and term polarity prediction. To enhance RRC performance, they introduced a post-training algorithm, which additionally pre-train the model on out-of-domain data from Amazon and Yelp review datasets, and additionally on MRC question answering dataset (Rajpurkar et al., 2016). These result in additional training set of $1,151,863$ for laptop domain, $2,677,025$ more examples for restaurant domain, and $87,599$ training examples from MRC dataset.

Karimi et al. (2020) proposed an approach based on conditional random field (CRF) (Lafferty et al., 2001), combined with BERT for aspect term extraction and term polarity prediction tasks. Two modules are employed for improving aspect term extraction and term polarity prediction of BERT model. First, a parallel approach is used which combines predictions for aspect term and polarity from last four layers of BERT in parallel. Moreover, a hierarchical aggregation module is also examined, where predictions of previous layers of BERT are fed into the next layer. Reddy et al. (2020) combines GLOVE pre-trained embedding (Pennington et al., 2014) with deep contextualized representation of BERT to enhance the representation of word vectors for predicting aspect term polarity. The proposed BERT-IL model predicts aspect term polarity by learning a similarity between GLOVE vector of aspect term and its contextualized representation extracted from BERT. First, the aspect term representations are extracted from multiple layers of BERT, and fed into a self-attention layer. Finally, it is further fine-tuned on ABSA task for performance improvement. In section 4, evaluation of our proposed generative language model are compared with the recent BERT-based models.

3 Model

This section describes aspect-based sentiment analysis task (ABSA), the proposed generative language model approach, details of the datasets, model training, and evaluation metrics.

3.1 Aspect Based Sentiment Analysis

Aspect-based sentiment analysis (ABSA) is similar to sentiment analysis, in the sense that the task is to predict the polarity of an entity in a sentence. However, it is different, since the goal is to predict fine-grained sentiment of multiple aspect terms and categories of an entity. The task was first introduced in Semantic Evaluation Challenge (SemEval14) (Pontiki et al., 2014). It was then extended in SemEval16 challenge (Pontiki et al., 2016). The challenges comprise of two domains, restaurant and laptop, where each domain spans over four sub-tasks (SB1-4).

Aspect Term Extraction (SB1) For a given review sentence, this sub-task is about predicting all aspects terms (word span) that opinions are expressed. It requires that all aspect terms to be predicted, including those which no opinion is expressed (neutral sentiment). This sub-task (AE) corresponds to sub-task 1 (SB1) - single sentence – slot 2 in SemEval16 challenge, named as opinion target expression (OTE) (Pontiki et al., 2016).

Aspect Term Polarity (SB2) For a given review sentence and an aspect term, the goal is to predict the polarity of the expressed opinion (`positive`, `negative`, `neutral`, `conflict`). This sub-task corresponds to SB1-Slot3 in SemEval16 challenge.

Aspect Category Detection (SB3) Given a set of pre-defined aspect categories (e.g. `PRICE`, `FOOD`, `SERVICE`, `AMBIENCE`, `ANECDOTE/MISCELLANEOUS`), the goal is to predict all categories that an opinion is expressed about. This sub-task corresponds to SB1-Slot1 (single-sentence) in SemEval16 challenge, where the category is defined as the pair of entity and attribute, e.g. `RESTAURANT#PRICE`, `FOOD#QUALITY`, `LAPTOP#GENERAL`, `LAPTOP#PRICE`. Please refer to Table 4 in the appendix for the full list of categories for laptop and restaurant domains.

Aspect Category Polarity (SB4) Given a review sentence and a category, the goal is to predict the sentiment of the category (`positive`, `negative`, `neutral`, `conflict`). This sub-task corresponds to SB1-Slot3 in SemEval16 (Pontiki et al., 2016).

3.2 Generative Language Modeling

ABSA task comprises of four sub-tasks: aspect term extraction, aspect category detection, and aspect term and category polarity predictions. The dominant approach for solving ABSA task is to train separate classifiers for each sub-task (Xu et al., 2019). In this paper, we propose to solve all sub-tasks using a single auto-regressive (generative) language model, either using single-task or joint-task training.

3.2.1 Language model

The goal of generative language modeling is to learn data distribution $p(x)$, where $x = (x_1, \ldots, x_n)$ is a sequence of n symbols. In order to model $p(x)$, the language model factorizes the distribution of a single sequence $p(x)$ using the chain rule of probability (Bengio et al., 2003), and training a neural network, which is parameterized by θ, by minimizing the negative log-likelihood,

$$p_\theta(x) = \prod_{t=0}^{n} p_\theta(x_t|x_{<t}) \quad (1)$$

$$\mathcal{L}_D = -\sum_{k=1}^{K}\sum_{t=1}^{n} \log p_\theta(x_t^k|x_{<t}^k) \quad (2)$$

During inference, the generative model sequentially generates tokens by conditioning on the input example x^k, and the past generated tokens.

3.3 ABSA task as generative language modeling

Each ABSA task training example, x^k, contains a sentence S^k, I pairs of aspect term and term polarity, and J pairs of aspect category and category polarity,

$$T^k = \{TP_i^k = (t_i^k, pt_i^k); i \in I\} \quad (3)$$

$$C^k = \{CP_j^k = (c_j^k, pc_j^k); j \in J\} \quad (4)$$

where t_i^k, pt_i^k, and TP_i^k are i-th aspect term, term polarity, and their pair. Moreover, c_j^k and pc_j^k, and CP_j^k are j-th aspect category, category polarity, and their pair of k-th sentence.

3.3.1 Single-Task Polarity Prediction

This task consists of predicting the polarity of aspect terms or aspect categories only (named as SB2 and SB4 in section 3.1). To generate polarity during the inference, the input to the generative language model (LM) comprises of k-th sentence and the corresponding aspect term or category,

$$pt_i^k = LM_{term}(S^k, t_i^k) \quad (5)$$

$$pc_j^k = LM_{category}(S^k, c_j^k) \quad (6)$$

where LM_{term} refers to a model that trained on aspect term dataset, and $LM_{category}$ refers to aspect category dataset, respectively. The details of training language model are described in section 3.3.3. Moreover, the details of input sequence formulation during training and inference are presented in Appendix A and Tables 3 and 5.

3.3.2 Joint and Multi-Task Prediction

This task includes generating pairs of aspect term and term polarity, or pairs of aspect category and their polarity. To jointly generate aspect terms and their polarities, the model input relies on the review sentence S^k only, and the model outputs all aspect term and polarity pairs in token-by-token (auto-regressive) generation,

$$T^k = LM_{term}(S^k) \quad (7)$$

$$C^k = LM_{category}(S^k) \quad (8)$$

where T^k is the set of aspect term and polarity pairs, Eq. (3), and C^k is the set of aspect category and polarity pairs, Eq. (4). The same method in joint-task prediction can be used to generate all pairs of aspect term and aspect category, i.e. multi-task prediction,

$$[T^k; C^k] = LM_{multi}(S^K) \quad (9)$$

In this case, during training, the model learns to generate I pairs of aspect term and J pairs of aspect category via language model training, Eq. (1).

3.3.3 Training

A training sequence for solving each sub-tasks (SB1-4) of section 3.1, consists of the review sentence, concatenated by the corresponding aspect term/category and its polarity. For example, in training LM_{term} for predicting aspect term polarity (Eq. 5) and joint-task prediction of aspect term and polarity (Eq. 7), the training sequence comprises of the review sentence concatenated by aspect terms and their polarities, $x^k = [S^k; T^k]$. Respectively, $x^k = [S^k; C^k]$ is used for training $LM_{category}$, as mentioned in Eq. (6) and (8). For more details on input sequence representation, see Appendix A, Tables 3 and 5.

In order to train LM_{term}, the model can be trained on different training sequences, where the review sentence S^k needs to only be concatenated with a single pair of aspect term and polarity. In this case, multiple training sequences are created for the k-th sentence, i.e. $\{x_i^k = [S^k; TP_i^k]; i \in I\}$. We will present An ablation study on these two methods of sequence creating for the language model training, and its effect on few-shot and full-shot performances, are presented in Appendix C and Figure 4.

3.4 Dataset

The proposed generative language model is evaluated on the two datasets proposed for ABSA task. SemEval14 challenge (Pontiki et al., 2014) consists of four sub-tasks as described in section 3.1. We

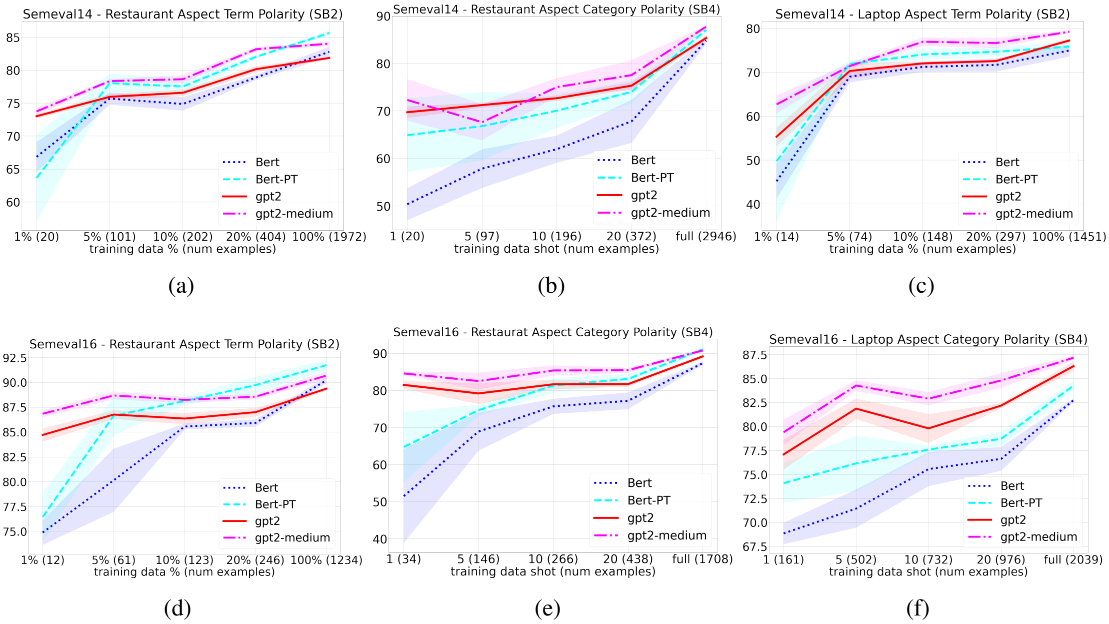


Figure 1: Single-Task polarity prediction (SB2 and SB4 sub-tasks), in few and full-shot settings. Note: 1-shot refers to one example per class, for aspect category, and 1% is percentage of training data for aspect term. Lines represents mean accuracy, and shaded area are standard deviation of experiments with 4 different random seeds. (best viewed in color)

Dataset	Domain	Train	Dev	Test
SemEval 14	Restaurant	3041	-	800
	Laptop	3045	-	800
SemEval 16	Restaurant	2000	-	676
	Laptop	2500	-	808
SST-2	Movie	66749	872	1821
SST-5	Movie	8544	1101	2210
OOS	Misc.	15100	3100	4500

Table 1: Dataset distribution

also evaluate the proposed model on task 5 of SemEval16 (Pontiki et al., 2016), which contains two sub-tasks for sentence and text level review data in multiple languages. In this paper, we only focus on the English language of sub-task 1 (sentence level) to be able to compare with the prior arts.

Moreover, we evaluate on Stanford Sentiment Treebank (SST) dataset (Socher et al., 2013) for binary (SST-2) and fine-grained (SST-5) sentiment classification of movie reviews domain. Since intent detection is a similar task to sentiment analysis, the evaluation is also performed on out-of-scope (OOS) intent detection dataset (Larson et al., 2019) which created for chatbot systems.

To evaluate the performance on few-shot setting, we sub-sample training set for aspect term and aspect category domains. For aspect term, the train set is randomly sub-sampled to the smaller sizes, $[1\%, 5\%, 10\%, 20\%]$. For example, 1% few-shot train set contains only about ≈ 20 sentences. For aspect category, since there is the predefined set of categories, we randomly sub-sample examples for each category, with different number of examples of $[1, 5, 10, 20]$.

The distribution of the train, dev and test splits for each domain are shown in Table 1. It is noteworthy that the previous baselines have created customized validation set from train set. Since no official validation set is released for SemEval14 and SemEval16, and in order to have a unified evaluation, we used the official `trial` set (part of train set) for validation, and exclude those examples from the train set. Moreover, prior works excluded examples with `conflict` polarity from their evaluations, since it is considered a difficult prediction task. However, for more accurate evaluation, these examples are retained in our evaluation.

3.5 Evaluation

Performance evaluation of aspect term polarity (SB2) and aspect category polarity (SB4) single-tasks in Eq. (5) and Eq. (6) are based on accuracy metric. It is measured by counting the number of aspect term and aspect category polarities which are correctly predicted. The evaluation of aspect term extraction (SB1) and aspect category detection (SB3) are measured by F1 metric (Pontiki et al., 2014) computed on the overlap of the ground-truth and generated sequences. The evaluation of SST-2, SST-5 and OOS datasets are measured by accuracy metric. On OOS dataset, full accuracy on in-domain and out-of-scope examples are measured.

Evaluation of joint and multi-task models in Eq. (7)(8)(9) are measured by joint accuracy. This means that for an example sentence S^k, if all the aspect term and term polarity predictions are correct, it is assumed as a correct prediction.

The restaurant domain contains both aspect term and aspect category annotations for SemEval14 and SemEval16. However, the laptop domain only contains aspect term annotation for SemEval14, and aspect category annotation for SemEval16. Therefore, single-task evaluation on laptop domain is constrained and multi-task prediction performance can only be evaluated on restaurant domain.

4 Experiments

In this section, the evaluation results are presented. The proposed generative language model is evaluated on five tasks. Single-task setting includes aspect term polarity and aspect category polarity prediction, Eq. (5)(6), for restaurant and laptop domains. Joint-task includes a) aspect term extraction and polarity Eq. (7) and b) aspect category detection and polarity Eq. (8). Finally, multi-task setting comprises all sub-tasks, i.e. aspect term extraction (SB1), aspect category detection (SB3), and their polarity predictions (SB2 and SB4), Eq. (9).

The evaluation of our proposed generative language model is compared with recent BERT-PT (Xu et al., 2019) model. We have reproduced results of BERT-PT on full-shot settings, since we include examples with `conflict` polarity. Other BERT-based models such as BERT-IL (Reddy et al., 2020) has not open-sourced code, and therefore they are not included in few-shot evaluation.

4.1 Single-Task Polarity evaluation

In this section, the proposed generative language model is evaluated on aspect term and aspect category polarity prediction for both restaurant and laptop domains. As shown in Figure 1, the proposed model, based on *GPT2-base*, outperforms BERT on few- and full-shot settings on all sub-tasks (SB2 and SB4) for SemEval14 and SemEval16. More importantly, GPT2 model has lower variance than BERT, especially in 1% or 1-shot setting.

Is it shown that BERT average performance drops by a large margin on low-resource regimes ($< 5\%$ or < 5 shot) and with increased variance, whereas our proposed generative model shows robust performance on few-shot setting with small variance. Compared to BERT-PT (Xu et al., 2019), which exploits additional pre-training on review data from Amazon and Yelp datasets, and using auxiliary tasks of MRC, generative model with more layers (*GPT2-medium*) and no additianl pre-training matches or outperforms BERT-PT average performance in few-shot setting with smaller variance. Interestingly, *GPT2-base* model (12 layers) outperforms BERT-PT average performance in some cases, including all 1% and 1-shot settings with reduced variance. For example, *GPT2-base* outperforms by a large margin, 16.75 points on average accuracy and reduces standard deviation by 8.8 points on 1%-shot setting of category polarity prediction in restaurant domain of SemEval16, Figure 1(e). Moreover, *GPT2-base* outperforms BERT-PT in all few- and full-shot settings on aspect category polarity prediction task (SB4) of restaurant domains in SemEval16 dataset, Figure 1(f).

Although *GPT2-medium* average performance mostly outperforms BERT-PT, there are some exceptions, such as Figure 1(a) for full-shot, Figure 1(c) for 5%-shot, Figure 1(d) for 20% and full-shot. On the other hand, BERT-PT has much larger variance and less robustness in all few- and full-shot settings. This is perhaps due to the use of out-of-domain data in additional pre-training of BERT-PT which results in higher variance, even than BERT baseline, when finetuned on few-shot downstream tasks. The goal of our proposed model is not to simply outperforms BERT-PT by additional pre-training, but to provide a robust model for few-shot setting.

More evaluation on sentiment polarity prediction on SST2, SST5 and OOS intent detection datasets are presented in Figure 2, Appendix G and Figure 8. They indicate that generative language model outperforms BERT-based classifier models. Overall, the results of single-task polarity prediction indicate that our proposed generative model based on language generation (uni-directional self-attention) have better performance than the discriminative models which uses BERT (bi-directional self-attention) as encoder.

4.2 Joint and Multi-Task evaluation

In this section, the proposed generative model is evaluated for joint and multi-task prediction. It includes solving two sub-tasks jointly, e.g. aspect term extraction and term polarity prediction, or aspect category detection and category polarity prediction, Eqs. (7)(8), or predicting all Eqs. (9). Since BERT and BERT-PT are single-task models, which required to use different heads for each sub-task,

Method	Training Task	Model	Restaurant		Laptop	
			Joint Accuracy	SB1 (F1)	Joint Accuracy	SB1 (F1)
Discriminative	Single (SB1)	MGAN	-	71.48	-	71.42
		BERT	-	74.1	-	79.28
		BERT-DK	-	77.02	-	83.55
		BERT-MRC	-	74.21	-	81.06
		BERT-PT	-	77.97	-	84.26
		BERT-PSUM	-	-	-	85.94
		BERT-HSUM	-	-	-	**86.09**
Generative	Joint (SB1&2)	GPT2 (base)	$56.47_{\pm 0.82}$	$77.59_{\pm 0.32}$	$50.65_{\pm 1.04}$	$72.61_{\pm 1.03}$
		GPT2 (medium)	$60.07_{\pm 0.52}$	**81.52**$_{\pm 0.8}$	$53.55_{\pm 0.43}$	$75.94_{\pm 0.17}$
	Multi (SB1-4)	GPT2 (base)	$49.84_{\pm 1.03}$	$77.92_{\pm 0.53}$	-	-
		GPT2 (medium)	$54.43_{\pm 0.47}$	**82.04**$_{\pm 0.21}$	-	-

Table 2: SemEval14 SB1 and SB2 sub-tasks for restaurant and laptop domains. Comparing joint and multi-task generative model with single-task BERT baselines for full-shot setting.

we can not directly compare our joint-task model with these baselines on join-accuracy metric. For example, BERT-PT uses groundtruth aspect term to evaluate on polarity prediction (SB2), which is not comparable to our joint-task model which generates aspect term and polarity jointly.

Results in Table 2 indicate that although generative model is trained in joint-task manner, for predicting aspect term extraction and term polarity, it still outperforms BERT-PT and other BERT baselines which are trained to solve single-task aspect term extraction only, on aspect term extraction (SB1) metric, in restaurant domain. However, in laptop domain, the generative model under-performs BERT-based models on aspect term extraction (SB1) metric, perhaps due to less training data in laptop domain for joint-task loss.

Aspect category sub-tasks improve aspect term extraction: In multi-task setting, where generative model is trained on all sub-tasks (SB1-4), the aspect term extraction (SB1) F1 metric is improved more, compared to when trained as a single-task model. This indicates that training the generative model using extra supervision (from aspect category) helps to extract multiple aspect terms in the review sentence more accurately.

Generative language modeling is better for multi-task learning: Evaluation results on SemEval14 restaurant domain are shown in Appendix B Table 6. Combined with the results from Table 2, it indicates that the proposed generative language model performs well on solving all sub-tasks (SB1-4) using language generation. For example, compared to joint-task setting (Table 2), aspect term extraction (SB1) F1 metric improves more for restaurant domain. Multi-task evaluation results on SemEval16 restaurant domain are shown in Appendix B Table 7 for reference.

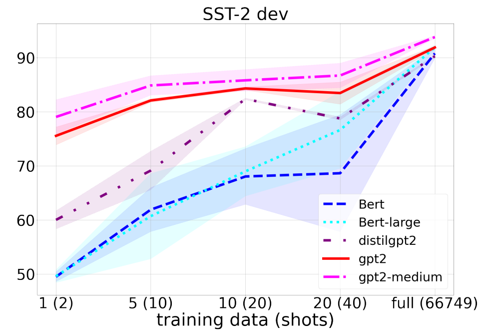
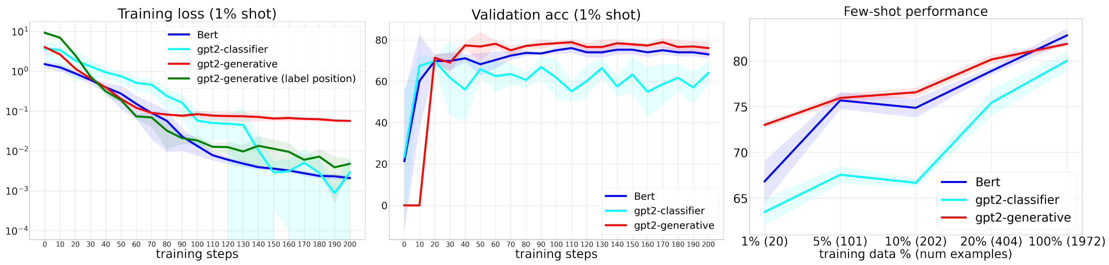


Figure 2: Few-shot evaluation on SST2 dev set. Note: 1-shot refers to one example per class. (best viewed in color)

4.3 Ablation

In this section, the ablation study of proposed generative language model is studied on two aspects. First, using the language model (GPT2) as a discriminative classifier vs. for language generation. Second, we study the training convergence of generative model with two discriminative baselines, i.e. BERT and GPT2 as classifier to better understand few-shot performance.

Generative vs. Discriminative training of unidirectional language model: To analyze the benefit of fine-tuning GPT2 using language modeling loss, we also fine-tune it as a classifier. In the latter case, a classification layer is added, which uses the output of the last token of the input sequence for polarity prediction. As shown in Figure 3(c), *GPT2-classifier* under-performs BERT, when only trained with discriminative loss. We conjecture that since GPT2 uses uni-directional self-attention (left-to-right), it captures less contextualized representation, compared to bidirectional self-attention in BERT. On the other hand, when fine-tuning GPT2 using generative loss (next word prediction), uni-directional self-attention learns a better representation, which improves few-shot performance. Ablation analysis on laptop domain and aspect category



(a) (b) (c)

Figure 3: Analysis of few-shot training convergence, evaluated on SemEval14 aspect term polarity prediction (SB2) on restaurant domain for 1% training data. GPT2-classifier model uses a classification layer on the output of last input token without using language modeling loss for training. *Note: Lines represents mean value, and shaded area are standard deviation of experiments with 4 random seeds.* (best viewed in color)

polarity predictions for both domains are shown in Appendix D and Figures 5 and 6.

GPT2 language model exploits more supervision than BERT in few-shot setting: To understand the training dynamics of generative language model and its relation to few-shot performance, we investigate the training convergence for GPT2, BERT, and *GPT2-classifier*. Results for SemEval14 restaurant aspect term polarity prediction are shown in Figure 3. It is indicated that BERT model converges faster than GPT2 in 1% few-shot settings, due to using a small classification head (fully-connected layer with 4 outputs) for the downstream task, which perhaps makes the model to overfits quickly to few-shot training data. On the other hand, GPT2 converges more slowly, perhaps due to using language modeling loss, i.e. cross-entropy loss across all tokens of the input sequence, and also using output layer with size of the vocabulary. However, the cross-entropy loss on the position corresponding to predicting label, *gpt2-generative (label position)*, converges faster than BERT, early in training, and the loss value is smaller than BERT between 40-90 steps, where the model has better validation accuracy than BERT. Later during the training, BERT training loss converges to smaller values, but its performance does not outperform GPT2. This is perhaps an evidence of BERT model overfitting due to using a small classification head which is specifically designed for the downstream task (4 output nodes).

Since the language modeling loss benefits GPT2 model to exploit more supervision during training (predicting words for all tokens in the input sequence), perhaps this helps the GPT2 to be less prune to overfitting, and outperforms BERT in few-shot setting. Additionally, reformulating the task as natural text might benefits GPT2 to infer the sentiment polarity easier than BERT. Overall, GPT2 validation and test accuracy achieves higher performance. Analysis of training convergence on other tasks and domains are presented in Appendix E, Figures 5 and 6.

We also investigates model parameter changes during finetuning by measuring the average of the absolute value of weights changes (aggregate shift) during training (more details are presented in Appendix F and Figure 7). It is shown that the aggregate shift of GPT2 parameters are $\approx 1e-4$ during training, while BERT aggregate shift reduces to $\approx 1e-7$, with same pattern observed for self-attention layers too. This indicates the benefits of language modeling loss, which gives GPT2 extra supervision to adapt to the few-shot data.

5 Conclusion

In this paper, we proposed to use a generative language model for aspect based sentiment analysis (ABSA). By reformulating the task as language generation, the model learns to predict aspects and their polarities via language generation. Evaluation results on single-task polarity prediction on few and full shot setting indicate that the proposed approach outperforms prior arts, which are based on discriminative classification using BERT as encoder, with higher average performance and lower variance. On join-task and multi-task settings, the proposed model shows better performance on single-task polarity prediction metrics. Additionally, evaluation results on coarse-grain (SST2), fine-grain (SST5) sentiment analysis datasets, and OOS intent detection dataset indicate the better and more robust few-shot performance of generative language model. Furthermore, qualitative analysis indicates that using language generation on multi-task setting improves the model prediction using supervision across aspect term and category.

6 Broader Impact

This work may have implications for the simplification of sentiment analysis using neural text generation. In the narrow sense, this work addresses aspect-based sentiment analysis. If so, the improvement of neural text generation systems and easier deployment would amplify both the positive and negative aspects of sentiment analysis. On the positive side, neural text generation models might play a role in automating user opinion mining, and thereby increasing efficiency of currently modular systems. On the negative side, it can dehumanize current systems, by automating systems towards multi-tasking, and reducing the level of human control on language generation. Moreover, this approach can introduce toxicity and biases into sentiment polarity predictions, such as gender, race, religious, and ethics (Kiritchenko and Mohammad, 2018; Park et al., 2018). This is due to biases which are learned during pretraining of neural text models on internet data (Sheng et al., 2019; Tan and Celis, 2019). These consequences are not specific to this work, but should be considered by the field of natural language processing more broadly.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Erik Cambria and Amir Hussain. 2012. Sentic computing. *marketing*, 59(2):557–577.

Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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

Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, and Xin Li. 2019. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4253–4260.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the*

8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language_understanding_paper.pdf*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Natesh Reddy, Pranaydeep Singh, and Muktabh Mayank Srivastava. 2020. Does bert understand sentiment? leveraging comparisons between contextual and non-contextual embeddings to improve aspect-based sentiment models. *arXiv preprint arXiv:2011.11673*.

Amani K Samha, Yuefeng Li, and Jinglan Zhang. 2014. Aspect-based opinion extraction from customer reviews. *arXiv preprint arXiv:1404.1982*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *arXiv preprint arXiv:1911.01485*.

Peter D Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

A Input Representation and Method Overview

As described in Section 3.3.3, a single training sequence consists of the concatenation of review sentence S^k with the corresponding aspect terms and their polarities $x^k = \left[S^k; T^k\right]$, or aspect categories and their polarities $x^k = \left[S^k; C^k\right]$.

A schematic overview of each segment is shown in Table 3 together with special tokens marking transition points. The generative language model is optimized by minimizing the negative likelihood over the joint sequence. The output state associated with each input token is used to predict the next token. During inference, for single task polarity prediction of each aspect term (sub-task SB1), the language model input comprises the review sentence concatenated by the corresponding aspect term. The the model generates a single token, which assumed as predicted polarity. Same method is used for sub-task SB4 for aspect category polarity prediction. For joint- and multi-task prediction, the input sequence contains only the review sentence. The language model then generates aspect terms and aspect categories along with their polarities in single toke-by-token generation, until the end-of-sentence special token is generated.

Examples of different input sequence formatting for different datasets evaluated in the paper are presented in Table 5. We are using identifiers to separate different segments of the input sequence. For example, to separate review sentence from aspect term, we introduced identifiers <|*review*|> and <|*term*|> to separate them. each segment also ends with an end-of-segment identifier, such as <|*endofreview*|> and <|*endofterm*|> identifiers. It is noteworthy that these identifiers are not special token, similar to BERT, which introduces new embeddings into vocabulary. We have noticed that defining identifiers as special token will decrease the performance of generative language model, perhaps due to introducing randomly-initialized embedding vectors into vocabulary, which requires more training data to finetune them. However, since GPT2 did not use special tokens during pretraining, using identifiers which are combination of pretrained vocabulary tokens and special characters, such as {<, |, ,|, >}, helps GPT2 to understand different segments in the input sequence, to infer the sentiment polarity more accurately.

B Multi-task prediction

In this section, evaluation results on SemEval 14 and SemEval16 restaurant domain are presented for multi-task learning using our proposed generative language model, based *GPT2-base* model, in Tables 6 and 7. For more details, please refer to section 4.2.

C Ablation: Model input sequence formatting

For a single review sentence with multiple aspect terms or categories, there are two ways to create input sequence for language model training, as described in section 3.3.3. First, the review sentence can be concatenated with each aspect terms separately (*GPT2-Split*), which results in better performance for few-shot setting (Figure 4) There are very few example in few-shot setting, such as 20 unique examples in 1% setting, and using split method increases training data and perhaps mitigates model over-fitting. However, when the review sentence is concatenated with all pairs of aspect terms or categories in a single sequence, performance is better for full-shot setting. There are few exceptions in Figure 4(a) for 1% and 5% shot settings. We observe that 1% few-shot contains 20, 14, 12 input sequences in Figure 4(a), (b), and (c), respectively, for the regular method. However, the split method increases input training sequences to 36, 23, 17. It means that when the number of training sequences are high enough, increasing number of training examples using split methods might deteriorates the few-shot performance, as shown in Figure 4(a). We guess that the better few-shot performance of the *GPT2-Split* method possibly depends on the number of unique training sequences when comparing to the regular method. In other words, the *GPT2-Split* methods might outperforms the regular method when the number of training sequences is very low.

D Ablation: Generative vs. Discriminative language model

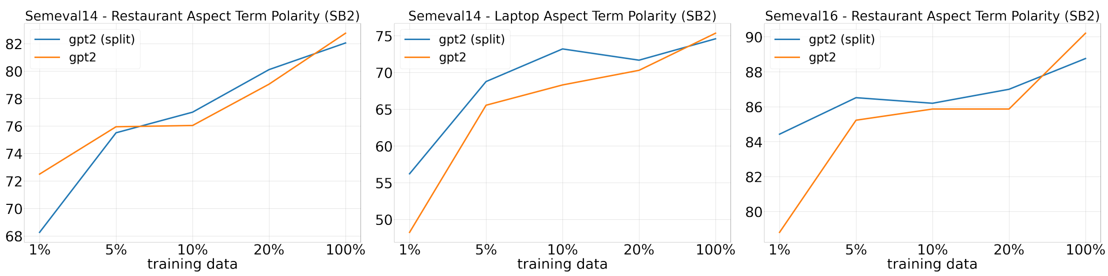
In this section, ablation analysis on using generative language model as a classifier are presented in Figures 5 and 6. It is shown that when fine-tuning *GPT2* model as a classifier on the downstream task using an classification layer, it under-performs BERT model on few and full-shot settings. For more details, please refer to section 4.3.

Sentence S^k	[review] *review sentence* [endofreview]
Aspect term T^k	[term] $term_1$ $polarity_1$, $term_2$ $polarity_2$, ... $term_I$ $polarity_I$ [endofterm]
Aspect category C^k	[category] $category_1$ $polarity_1$, $category_2$ $polarity_2$, ... $category_J$ $polarity_J$ [endofcategory]
Aspect term single and joint task training sequence (LM_{term})	[review] *review sentence* [endofreview] [term] $term_1$ $polarity_1$, ... [endofterm]
Aspect category single and joint task training sequence ($LM_{category}$)	[review] *review sentence* [endofreview] [category] $category_1$ $polarity_1$, ... [endofcategory]
Multi-task training sequence (LM_{multi})	[review] *review sentence* [endofreview] [term] $term_1$ $polarity_1$, ... [endofterm] [category] $category_1$ $polarity_1$, ... [endofcategory]

Table 3: A schematic representation of the different components of inputs/outputs in aspect-based sentiment analysis. When training generative language model, these are concatenated together into a single sequence, as shown in last three rows.

Dataset	Domain	Aspect Category	
		Entity	Attribute
SemEval 14	Restaurant	ambience, anecdotes miscellaneous, food, price, service	N/A
	Laptop	N/A	N/A
SemEval 16	Restaurant	ambience, drinks, food, location, restaurant, service	general, price, style, quality
	Laptop	battery, company, cpu, display, fans cooling, graphics, hard disc, hardware, keyboard, laptop, memory, motherboard, mouse, multimedia devices, optical drives, os, ports, power supply, shipping, software, support, warranty	miscellaneous, operation performance, quality, general, design features, usability, connectivity, portability, price

Table 4: Ascpet category definition for SemEval14 and SemEval16 datasets. In Semeval14, each unique aspect category is defined as entity. For SemEval16, aspect category is defined as combination of entity and attribute. Laptop domain does not have annotation in SemEval14 dataset.



(a) (b) (c)

Figure 4: Ablation analysis on model input sequence formatting. *GPT2 (split)* means review sentence is concatenated with each aspect terms separately. (best viewed in color)

E Ablation: Training convergence

In this section, training convergence of GPT2 model is compared with BERT and GPT2-classifier model in varios tasks of aspect-based sentiment analysis. As shown in Figures 5 and 6, GPT2 achieves higher validation accuracy, when its training losses, standard language modeling and loss corresponding to label position, have higher value

Dataset	Task	Type	Input sequence	
			train	inference
SemEval14	Single task	aspect term polarity prediction	<|review|> once we sailed, the top-notch food and live entertainment sold us on a unforgettable evening. <|endofreview|> <|term|> food positive , live entertainment positive <|endofterm|>	<|review|> once we sailed, the top-notch food and live entertainment sold us on a unforgettable evening. <|endofreview|> <|term|> food
SemEval14	Joint task	aspect term	<|review|> once we sailed, the top-notch food and live entertainment sold us on a unforgettable evening. <|endofreview|> <|term|> food positive , live entertainment positive <|endofterm|>	<|review|> once we sailed, the top-notch food and live entertainment sold us on a unforgettable evening. <|endofreview|>
SemEval14	Multi-task	aspect term & aspect category	<|review|> the service was attentive without being overbearing and each dish we tried was wonderful from the spring rolls to the cod with pineapple tempura. <|endofreview|> <|term|> service positive , dish positive , spring rolls positive , cod with pineapple tempura positive <|endofterm|> <|category|> food positive , service positive <|endofcategory|>	<|review|> the service was attentive without being overbearing and each dish we tried was wonderful from the spring rolls to the cod with pineapple tempura. <|endofreview|>
SST-2	Single-task	polarity prediction	<|review|> does n't try to surprise us with plot twists , but rather seems to enjoy its own transparency <|endofreview|> <|sentiment|> positive <|endofsentiment|>	<|review|> does n't try to surprise us with plot twists , but rather seems to enjoy its own transparency <|endofreview|> <|sentiment|>
SST-5	Single-task	polarity prediction	<|review|> it 's a lovely film with lovely performances by buy and accorsi . <|endofreview|> <|sentiment|> somewhat positive <|endofsentiment|>	<|review|> it 's a lovely film with lovely performances by buy and accorsi . <|endofreview|> <|sentiment|>
OOS	Single-task	intent prediction	<|user|> how would you say fly in italian <|endofuser|> <|intent|> translate <|endofintent|>	<|user|> how would you say fly in italian <|endofuser|> <|intent|>

Table 5: Examples of input sequence during training and inference of generative language model for different datasets.

Shot	Layers	Joint Accuracy	Term		Category	
			SB1 (F1)	SB2 (Acc)	SB3 (F1)	SB4 (Acc)
1%	12	20.75	39.26	19.69	62.82	43.4
	24	20.62	37.87	18.99	61.79	41.51
5%	12	31	44.35	32.38	74.46	56.51
	24	34.87	60.4	35.18	75.39	59.06
10%	12	38.37	62.47	35.98	77.43	61.32
	24	41.75	65.9	40.06	79.27	62.92
20%	12	42.88	66.82	39.91	79.39	62.36
	24	45	72.73	45.31	80.79	65.28
100%	12	51.63	77.43	49.71	85.34	70.57
	24	55.62	81.53	57.92	82.4	70.38

Table 6: Multi-task evaluation on SemEval14 restaurant domain (SB1-4) on few-shot settings using generative language model (GPT2).

Shot	Layers	Joint Accuracy	Term		Category	
			SB1 (F1)	SB2 (Acc)	SB3 (F1)	SB4 (Acc)
1%	12	11.6	28.68	13.38	46.36	38.31
	24	9.04	24.87	11.36	44.32	35.63
5%	12	18.43	33.81	16.74	56.85	50.06
	24	20.48	34.99	18.88	61.09	54.66
10%	12	21.16	33.48	16.74	63.11	50.45
	24	22.18	37.13	19.64	67.12	55.43
20%	12	25.77	37.74	20.63	69.39	62.07
	24	26.96	40.6	22.15	72.9	65.39
100%	12	32.42	48.48	27.67	76.51	66.41
	24	43	50.27	30.15	76.78	69.6

Table 7: Multi-task evaluation on SemEval16 restaurant domain (SB1-4) on few-shot settings using generative language model (GPT2).

than BERT and GPT2-classifier. This indicates that perhaps BERT and GPT2-classifier overfitted to the few-shot training data. On the other hand, GPT2 language model achieves more supervision via standard language modeling loss, which results in higher training loss, but better validation performance.

F Ablation: Model parameters change during training

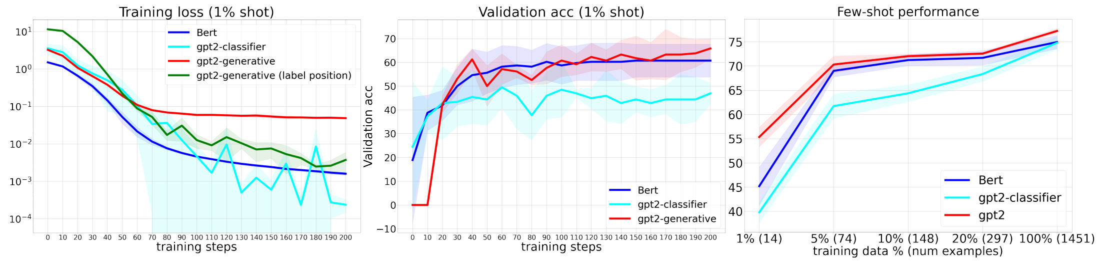
In order to understand model behavior during training on few-shot data, we study the change in parameters value of GPT2 and BERT models during training on 1% few-shot data. The aggregate shift is computed by the mean value of the absolute value of weight change for each parameter $|w_{i+1} - w_i|$. The comparison between GPT2 and BERT model when trained on 1% few shot data of SemEval14 restaurant domain are shown in Figure 7

G Ablation: Other Sentiment Analysis Tasks

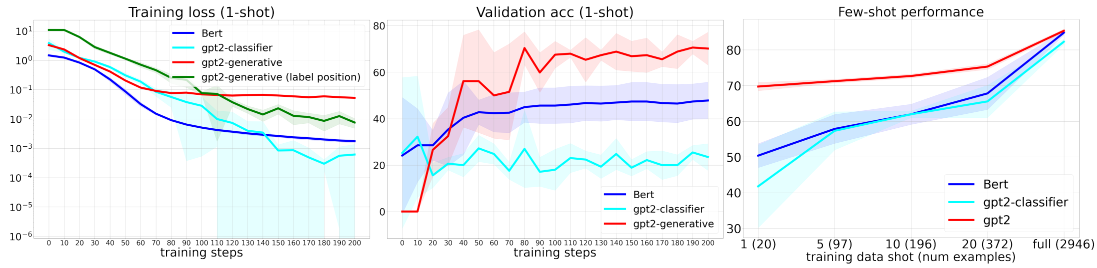
In order to extend the investigate the performance of our proposed generative language model to other sentiment analysis tasks, we also evaluate few-shot performance on SST-5 sentiment analysis dataset (Socher et al., 2013) (binary and fine-grained sentiment classification), and OOS (Larson et al., 2019) intent detection dataset. The results are shown in Figure 8, which indicate the superiority of generative model (GPT2) over discriminative BERT. On intent detection, Figure 8(c), GPT2 also outperforms TOD-BERT (Wu et al., 2020) which exploits extra pretraining on dialogue datasets to increase its few-shot performance.

H Qualitative Analysis

As described in section 4.2 and Table 2, aspect term extraction on restaurant domain (SemEval14) is improved in multi-task learning. To better understand model behavior, some examples are shown in Table 8. Using aspect category as supervision in multi-task learning helps the model to more accurately generates the aspect terms, reduces false positive aspect terms and wrong polarity predictions. Moreover, multi-tasking helps to better predict category polarity, using supervision from aspect term during training. Some examples of wrong prediction are shown in Table 9. It indicates that when there are negative or conflict polarity, the model struggles to correctly predict everything correctly. This often happens when there are opposite opinions for different aspect terms or categories.



(a) SemEval14 Laptop Aspect Term Polarity (SB2)

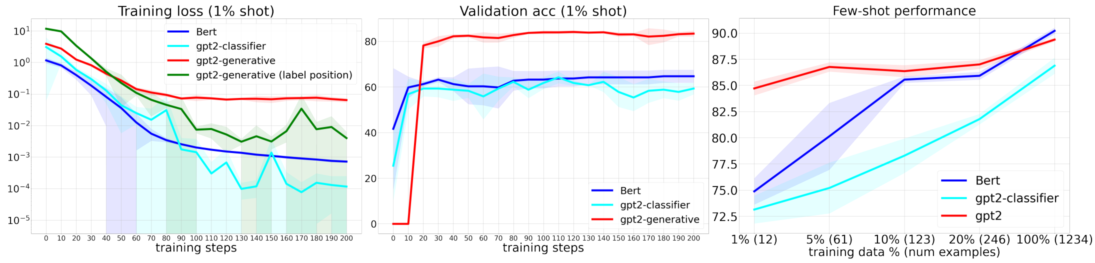


(b) SemEval14 Restaurant Aspect Category Polarity (SB4)

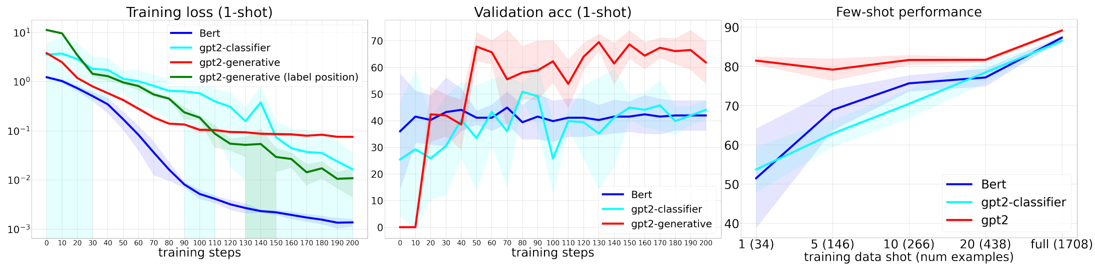
Figure 5: Analysis of few-shot training convergence, evaluated on SemEval14 for 1% and 1-shot training data, and few-shot performance on all settings (right). GPT2-classifier model uses a classification layer on the output of last input token without using language modeling loss for training. (best viewed in color)

<table>
<tr><th>Sentence</th><th>Task</th><th>Model Output</th></tr>
<tr><td rowspan="4">the sangria's - watered down.</td><td>aspect term</td><td><|term|> sangria negative</td></tr>
<tr><td>aspect category</td><td><|category|> food neutral</td></tr>
<tr><td>aspect term & category</td><td><|term|> sangria negative <|category|> food negative</td></tr>
<tr><td>groundtruth</td><td><|term|> sangria negative <|category|> food negative</td></tr>
<tr><td rowspan="4">everyone who works there (the host, the bartender, the servers) is so helpful.</td><td>aspect term</td><td>host positive, bartender neutral, servers positive</td></tr>
<tr><td>aspect category</td><td><|category|> service positive</td></tr>
<tr><td>aspect term & category</td><td><|term|> host positive, bartender positive, servers positive <|category|> service positive</td></tr>
<tr><td>groundtruth</td><td><|term|> bartender positive, host positive, servers positive <|category|> service positive</td></tr>
<tr><td rowspan="3">in mi burrito, here was nothing but dark chicken that had that cooked last week and just warmed up in a microwave taste.</td><td>aspect term</td><td><|term|> dark chicken negative, microwave taste neutral</td></tr>
<tr><td>aspect category</td><td><|category|> food negative</td></tr>
<tr><td>aspect term & category</td><td><|term|> chicken negative, taste negative <|category|> food negative</td></tr>
<tr><td rowspan="4">if you like seafood and or greek food you will love this place though it is not limited to just these things.</td><td>aspect term</td><td><|term|> seafood positive, greek food positive, place negative</td></tr>
<tr><td>aspect category</td><td><|category|> food positive</td></tr>
<tr><td>aspect term & category</td><td><|term|> seafood positive, greek food positive <|category|> food positive</td></tr>
<tr><td>groundtruth</td><td><|term|> greek food positive, seafood positive <|category|> food positive</td></tr>
</table>

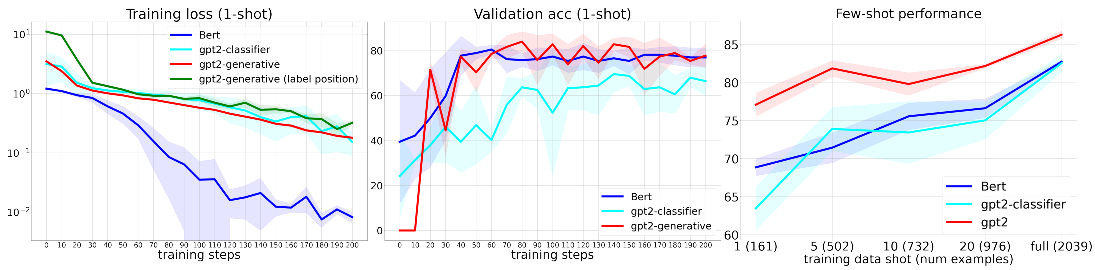
Table 8: Examples of correct predictions in multi-task learning.



(a) SemEval16 Restaurant Aspect Term Polarity (SB2)

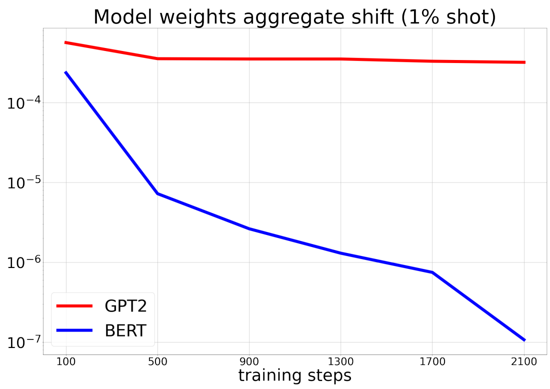


(b) SemEval16 Restaurant Aspect Category Polarity (SB4)

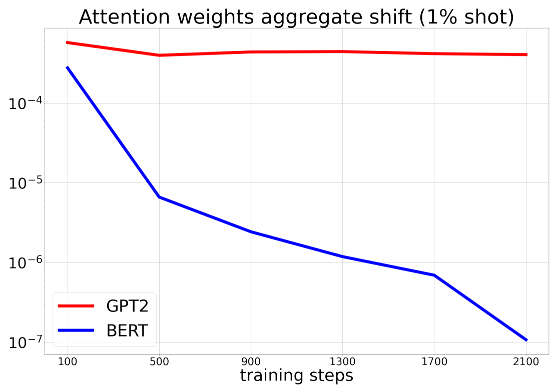


(c) SemEval16 Laptop Aspect Category Polarity (SB4)

Figure 6: Analysis of few-shot training convergence, evaluated on SemEval16 for 1% and 1-shot training data, and few-shot performance on all settings (right). GPT2-classifier model uses a classification layer on the output of last input token without using language modeling loss for training. (best viewed in color)

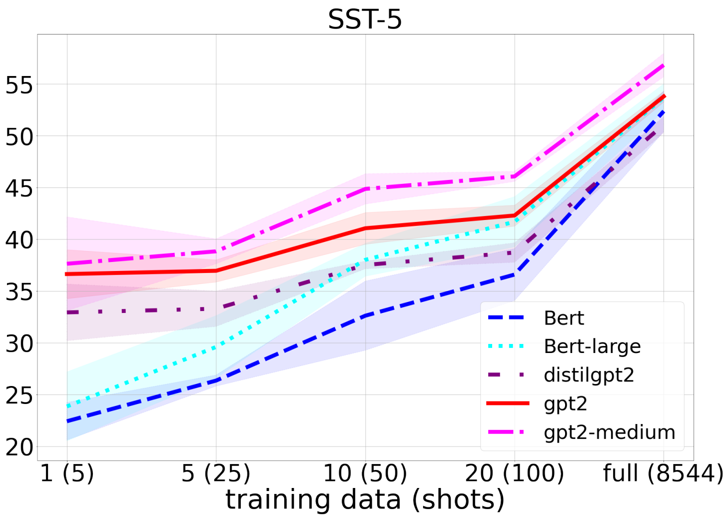


(a)

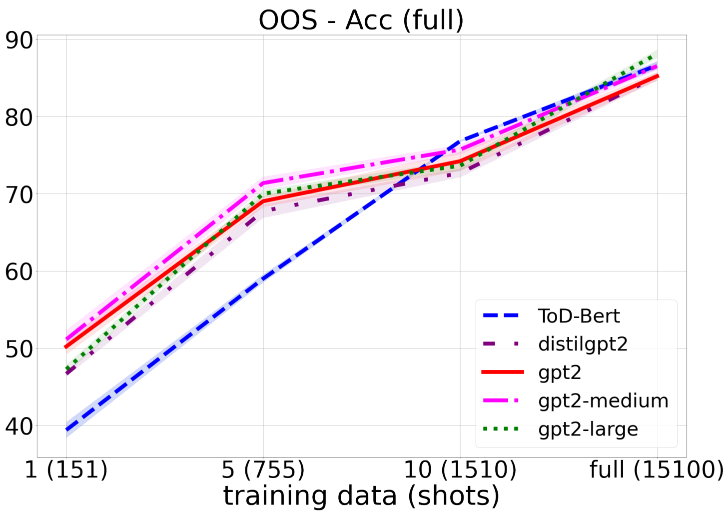


(b)

Figure 7: Analysis of few-shot performance through model parameter changes during training on 1% few-shot data on SemEval14 restaurant aspect term polarity (SB2) prediction task.



(a)



(b)

Figure 8: Few-shot evaluation of GPT2 and BERT models on SST5 and OOS intent detection datasets. Note: 1-shot refers to one example per class. (best viewed in color)

<table>
<tr><th>Sentence</th><th>Task</th><th>Model Output</th></tr>
<tr><td rowspan="4">certainly not the best sushi in new york, however, it is always fresh, and the place is very clean, sterile.</td><td>aspect term</td><td><|term|> sushi negative, place positive</td></tr>
<tr><td>aspect category</td><td><|category|> ambience positive, food positive</td></tr>
<tr><td>aspect term & category</td><td><|term|> sushi positive, place positive <|category|> food positive, ambience positive</td></tr>
<tr><td>groundtruth</td><td><|term|> place positive, sushi conflict <|category|> ambience positive, food conflict</td></tr>
<tr><td rowspan="4">while there's a decent menu, it shouldn't take ten minutes to get your drinks and 45 for a dessert pizza.</td><td>aspect term</td><td>menu positive, drinks positive, dessert pizza positive</td></tr>
<tr><td>aspect category</td><td>food conflict</td></tr>
<tr><td>aspect term & category</td><td><|term|> menu positive, drinks positive, dessert pizza positive <|category|> food positive</td></tr>
<tr><td>groundtruth</td><td><|term|> dessert pizza neutral, drinks neutral, menu positive <|category|> food positive, service negative</td></tr>
<tr><td rowspan="4">the portions of the food that came out were mediocre.</td><td>aspect term</td><td>portions negative, food neutral</td></tr>
<tr><td>aspect category</td><td>food negative</td></tr>
<tr><td>aspect term & category</td><td><|term|> portions negative, food negative <|category|> food negative</td></tr>
<tr><td>groundtruth</td><td><|term|> portions of the food neutral <|category|> food neutral</td></tr>
</table>

Table 9: Examples of wrong prediction for joint and multi-task generative language model.