An Adaptive Layer to Leverage Both Domain and Task Specific Information from Scarce Data

Gaël Guibon[1,2], **Matthieu Labeau**[1], **Luce Lefeuvre**[2], **Chloé Clavel**[1]

[1] LTCI, Télécom-Paris, Institut Polytechnique de Paris
[2] Direction Technologies, Innovation & Projets Groupe, SNCF
gael.guibon@gmail.com, {gael.guibon,matthieu.labeau,chloe.clavel}@telecom-paris.fr, luce.lefeuvre@sncf.fr

Abstract

Many companies make use of customer service chats to help the customer and try to solve their problem. However, customer service data is confidential and as such, cannot easily be shared in the research community. This also implies that these data are rarely labeled, making it difficult to take advantage of it with machine learning methods. In this paper we present the first work on a customer's problem status prediction and identification of problematic conversations. Given very small subsets of labeled textual conversations and unlabeled ones, we propose a semi-supervised framework dedicated to customer service data leveraging speaker role information to adapt the model to the domain and the task using a two-step process. Our framework, Task-Adaptive Fine-tuning, goes from predicting customer satisfaction to identifying the status of the customer's problem, with the latter being the main objective of the multi-task setting. It outperforms recent inductive semi-supervised approaches on this novel task while only considering a relatively low number of parameters to train on during the final target task. We believe it can not only serve models dedicated to customer service but also to any other application making use of confidential conversational data where labeled sets are rare. Source code is available at https://github.com/gguibon/taft

Introduction

Customer service is a crucial matter for companies (customer satisfaction, *etc.*) and academics (better understanding of dyadic conversations); however, customer service software clients require better handling. Indeed, customer service holds many obstacles, such as non publicly available data, huge annotation cost, confidentiality issues on several levels, and so on. This also implies multiple levels of analysis: conversation level, and utterance level. While this dual level of analysis has only been recently tackled in the literature (Yang et al. 2022), a conversation-level label has never been considered as a prediction target in this context. Studies focus mainly on predicting utterance-level labels (Zadeh et al. 2018a,b; Hazarika et al. 2018; Majumder et al. 2019). In this work, we consider both levels of conversational analysis, as our main purpose is to detect problematic conversations. To do so, we place ourselves in the use-case of a company with customer service chats, which implies that available data is scarce, confidential, and thus difficult to label – not to mention, costly. In this context, we investigate the best way to adapt an existing model to two multi-class classification tasks: predict the customer satisfaction, and predict the status of the customer problem, which are both conversation-level labels. For the former, we consider a representation of customer satisfaction as discrete polarity labels. For the latter, we consider the status of the problem as 5 possible values: aborted, problem solved, solution to be tested, problem out-of-scope or no solution available. As far as we know, this work is the first on customer problem status identification, even more in the context of customer service. We make the hypothesis that automatically predicting complex labels such as the status of a problem is feasible, even with a very small quantity of data.

In order to study this use-case, we use confidential customer service data that we annotated. Indeed, as far as we know, there is no publicly available dataset which contains a two-level annotation scheme in customer service chats, and no dataset containing status of problem resolution as labels. A real dedicated dataset is thus mandatory, even though it is confidential. The context of customer service implies that only a few conversations are available, and even fewer could be labeled, due to confidentiality. We decide to tackle this joint issue by using semi-supervised learning. The most fitting existing method in the literature seems to be domain-adaptive pre-training (Konlea and Jannidisa 2020; Wu et al. 2021), which consists in adapting an existing pre-trained language model to a new domain by continuing training with the Masked Language Model objective on new data. However, we suspect our data to be too scarce for this, and we propose a method to apply semi-supervised learning to adapt an existing, pre-trained language model by using a substitute, simple task, which does not imply any additional labeling cost. In the context of customer service, we choose to use *speaker role information*, which is almost always available. By doing so, we want to verify that trying to predict simple information already embedded in the structure of the data is a reliable way to replace predicting missing words, *i.e* masked language modelling. Secondly, in order to tackle these two classification tasks, we experiment with multi-task learning. Our approach lies on using a supplementary *adaptive* layer, which we will use to transition from our *adaptive task* of

identifying the speaker role, to our *final tasks*, giving overall a more malleable model and a faster fine-tuning.

Our contribution is three-fold. First, we present the first work on problem status identification for customer service conversations. Second, we apply a novel semi-supervised learning approach, leveraging available information, embedded in the structure of the data, to fine tune an existing pre-trained language model. Third, we apply a multi-task learning objective to better solve our two different, but related final tasks: predict customer satisfaction and predict the status of the customer problem. We believe this work can be useful for many real use-cases in customer service, especially when labels are complex and data is scarce. Finally, our approach is not restricted to predict satisfaction or problem related labels, but can be applied to other kinds of labels in a similar context.

Related Work

Customer Satisfaction and Customer Service. While a recent work looks at customer satisfaction through customer engagement on tweets (Liu, Shin, and Burns 2021), another recent approach represent the customer satisfaction as a Net Promoter Score (Piris and Gay 2021) (NPS) in the Orange company's Customer Service contact center logs using customer surveys on $\sim$80,000 conversations to predict the NPS score. They obtained a macro f1-score of 53.8 using a CNN on a binary classification task and 48.3 on a 3-label classification task. They showed it is best to take into account both customer and agent turns than to isolate them (Auguste et al. 2019). Other works used a mix of regular expression, semantics, sentence similarity and other approaches to consider the customer satisfaction through question answering (Gong et al. 2019). A more classical approach consisted in using VADER (Hutto and Gilbert 2014) to extract the sentiment score on 170k Samsung Chat Logs in USA, and then applying SVM, logistic regression and random forest to predict the customer's satisfaction (Park et al. 2015). They used a 5-scale satisfaction and obtained a high accuracy (0.84) but a low f1-score (0.39). Following this work and still using Samsung chat logs, satisfaction prediction incorporating time gaps in LSTM achieved 0.79 in f1-score. They used two sets, a labeled one from surveys and another set which was annotated by three humans with a kappa of 0.51. They also showed surveys often possess a positivity bias (Park, Cha, and Rhim 2018). Related spoken conversations were used considering 3 polarity values and turn-taking features extraction with lexical and prosodic feature sets, achieving a F1-score of 0.85 (Chowdhury et al. 2016), while other work identified turn-level and call-level estimations with joint biLSTM-RNN with a macro f1-score of 0.710 (Ando et al. 2017).

Speaker Role in Dialogue. Speaker identification has been used in dialog act detection in conversations, as a one-hot encoded feature (Bothe et al. 2018), or a sequence of identifiers (Shang et al. 2020), for instance. However, speaker role identification has mainly been seen has a task to solve in call centers conversations, with two roles (client or agent) on confidential data, with results achieving 93% in accuracy (Lavalley et al. 2010; Clavel et al. 2013). These works considered linguistic patterns, but more recent ones tried to integrate different speaker roles (receiver or sender) in multi-turn dialogue representation (Liu et al. 2021a) or tackled speaker role identification (air-traffic controller or pilot) using grammar rules in speech (Prasad et al. 2021). In our work, we consider the common customer–agent speaker roles and integrate this task in our approach to better detect problematic conversations and problem resolution status.

Problem Resolution in Customer Service. While problem detection in interactions has been done using handcrafted linguistic rules (Maslowski, Lagarde, and Clavel 2017), only a few recent works try to predict problem resolution, and none of them tried to predict it directly. Jain (2021) used information from the status of the problem resolution to check its correlation with the polarity of customer sentiment; however, they did not try to predict this status directly. Chen et al. (2021) created a dataset of customer service conversations and predicted the intent, considering constraints from the agent guidelines. While this dataset contains information about product flows, they do not try to predict it. Finally, customer feedback on emotional support conversations has been done, on related customer issues. While these issues are not of the same nature as the ones we study in this paper, the context is related. Once again, they do not try to predict the status of the problem resolution, nor do they try to predict the customer's feedback (Liu et al. 2021b). Contrarily to these previous works, we want to predict the status of the problem resolution, and to the best of our knowledge, we are the first to tackle this task directly.

Inductive Semi-Supervised Learning. In our work, we take advantage of semi-supervised learning for transfer learning; also referred to as inductive semi-supervised learning (Shi et al. 2009; Zhou et al. 2018). Among recent inductive semi-supervised learning approaches, there exist a few which tackle sentiment analysis with weak supervision on customer chat: Jain (2021) used weak labels to predict on unlabeled data, using lexical feature based classifiers, before fine-tuning a RoBERTa model (Liu et al. 2019) on it. Other recent inductive semi-supervised learning approaches consider either a task-adaptive pre-training (Howard and Ruder 2018) with discriminative fine-tuning, *i.e.* a different learning rate for each layer, or a domain-adaptive pre-training (Konlea and Jannidisa 2020; Wu et al. 2021). These methods have been shown to be really effective to adapt language models to a specific task before fine-tuning (Gururangan et al. 2020). In this work, considering the small quantity of data we have access to, we choose not to use adaptive pre-training, but rather use the simple task of speaker role identification to improve the detection of problematic conversations and the new task of problem resolution status detection in customer service, which can be viewed as *adaptive fine-tuning*.

Pruksachatkun et al. (2020) also defined an intermediate fine-tuning, using complex intermediate tasks. Contrarily to their conclusion, we show that in customer service context, a simple intermediate task can be sufficient to yield better results. On the other hand, dummy tasks, along with synthetic data that mimics the original, have been used in the intermediate fine-tuning framework (Chang and Lu 2021).

Notations

Our dataset $\mathcal{D}$ is comprised of conversations $(C_1, C_2, \ldots, C_{|D|})$, which are in turn made of utterances: $C_i = (u_1, u_2, \ldots, u_{|C_i|})$. Each utterance possesses a speaker role label a indicating whether it is from a customer or an operator (which is obtained automatically and does not need to be manually labeled). Utterances are sequences of words $u_j = (w_1^j, w_2^j, \ldots, w_{|u_j|}^j)$ concatenated using the separator variable <s/>. Thus, we can represent a conversation as a flattened list of words across utterances: $C_i = (w_1^1, w_2^1, \ldots, w_{|u_1|}^1, \texttt{<s/>}, w_1^2, \ldots, w_{u_{|C_i|}}^{|C_i|})$. Our dataset is divided in two parts: first, a subset where conversations are labeled with problem status p and satisfaction s. Hence, in this labeled subset $\mathcal{D}_L$, each conversation comes with these two labels: $\mathcal{D}_L = ((C_1, y_1^p, y_1^s), \ldots, (C_{|D_L|}, y_{|D_L|}^p, y_{|D_L|}^s))$. The second subset $\mathcal{D}_U$, unlabeled, is taken directly at the utterance level, with each utterance being associated to their speaker role label: $\mathcal{D}_U = ((u_1, y_1^a), (u_2, y_2^a), \ldots, (u_{|D_u|}, y_{|D_u|}^a))$

Tasks and Labels Definition

We will now explicit our 3 different labels and the tasks they are associated to.

Conversation-level tasks. At the conversation-level, we want to predict the status of the customer problem p, which is one of 5 distinct labels from which the last 4 are used for the request's status classification:

- **Aborted**: the conversation is cut short (*thus excluded from the prediction task*).
- **Solved**: the problem is solved. The operator gave the expected information or solution to the customer.
- **Test_required**: the operator gave steps to follow later on. The customer needs to try this procedure.
- **Out-of-scope**: the customer's issue cannot be tackled by the operator. The latter can either redirect the customer to another service, or simply indicate that this type of request cannot be tackled.
- **No solution**: No solution is provided to the customer by the operator, which cannot even redirect the customer to another service.

We also predict the customer satisfaction s, which simply consists in 3 discrete polarity values (-1, 0 and 1).

Utterance-level task. At the utterance-level, we will use speaker role identification as a binary classification task between two roles: *customer* and *operator*.

Methodology

Our goal, in the context of customer service, is to work on two target tasks: predict the customer satisfaction and identify the status of their problem. Given the small quantity of labeled data we have access to, we propose a dedicated approach for semi-supervised learning for customer service conversations, making use of our unlabeled data, but also the structural information that is tied to it – in our case, the speaker role labeling of the utterances.

Our method can be summarized by the following steps:

- (1) pre-train a language model on a general domain,
- (2) add a supplementary *adaptive* layer to this model,
- (3) train the resulting model on a simple task – predict the labels a, using the data $\mathcal{D}_U$,
- (4) freeze the initial model and fine-tune the adaptive layer on the target task – predict the labels p and s, using the data $\mathcal{D}_L$.

It can be seen as a mix between adding an intermediate fine-tuning step using an appropriate secondary task (3), and adding a supplementary *adaptive* layer (2), which would need a supplementary pre-training step, with data from the appropriate domain. Hence, we name our model Task-Adaptive Fine-Tuning (TAFT). As a supplementary help against our lack of data, we adopt a multi-task approach: we simultaneously fine-tune our adaptive layer in the last step on both our target tasks. As both tasks are correlated, we expect this to improve results.

Figure 1 shows our TAFT approach in detail. The top half of the figure represents step (3), in which we exploit the speaker role identification task for simple intermediate fine-tuning; indeed, this information is both always available in customer service written interactions and anonymous, which means customer service data almost always possess speaker role labels. This step can be seen as both *fine-tuning the initial model* and *pre-training the adaptive layer*, adapting it to the domain.The bottom half shows step (4): once this step is over, we keep the initial model, now adapted to the new domain, unchanged, by freezing the parameters, and apply the final fine-tuning on the adaptive layer with problem status identification and customer satisfaction detection tasks.

Using an Adaptive Layer . One of the key points of our method is the adaptive layer whose closest equivalent work is the GAFL from computer vision to pre-filter image frequencies (Shipitsin, Bespalov, and Dylov 2022). The purpose of the adaptive layer is to serve as an intermediate, lighter encoder, that will retain information about how to use the output of the initial language model, but is separated from it, which allows it to be fine-tuned independently afterwards. Our reasoning is that the smaller size of the layer, relatively to the initial model, will make fine-tuning easier – which is important when data is scarce – with the added benefit of having less parameters to train and a faster convergence.

Intermediate Fine-tuning on simple task. Our method employs speaker role identification as intermediate fine-tuning task; previously, this information has been used through dedicated embeddings (Yuji, Tetsuya, and Yoshi 2022), or as additional parts of the model (Ghosal et al. 2019; Bao et al. 2022). Our reasons are multiple: while these labels are easily available and require no manual annotation, they contain information about the data and its structure. But it is especially important to note that this is a simple task, correlated to the target tasks, by opposition with (Chang and Lu 2021): while recent work (Gururangan et al. 2020) suggests to use a complex intermediate task, we make the hypothesis something easy may allow the model to extract information more reliably, given that here again, the quantity of data

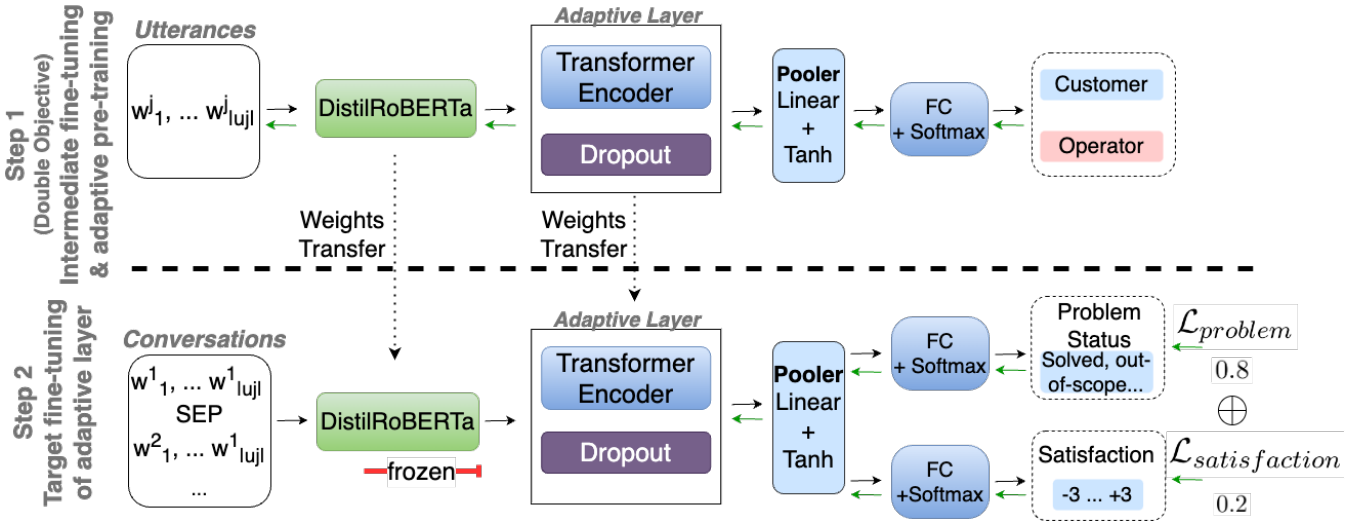

Figure 1: Representation of our approach in two main steps. From left to right, starting with a Pretrained Language Model and ending with a multitask fine-tuning for identifying the status of the customer's problem and predicting the customer satisfaction. The right-to-left arrows represent the backpropagation.

available is rather small. We also hypothesize that the initial language model will not adapt too much to a task this simple, which seems confirmed by how fast it converges to a very high classification score (even without an adaptive layer).

Multi-Task Fine-Tuning of the Adaptive Layer. Since the labels of both tasks have high correlation given their Jaccard similarity (Appendix A), we experiment with fine-tuning with a multi-task classification objective, where we choose the primary task to be the prediction of the problem status p and the secondary task to be the prediction of the customer satisfaction s, which, normally easier, should help reinforce the performance of the primary task. We use a cross entropy loss $\mathcal{L}(\hat{y}, y) = -\sum_k^K y^{(k)} \log \hat{y}^{(k)}$ for each task and we combine them in a weighted sum:

$$\mathcal{L} = 0.8 \times \mathcal{L}_p + 0.2 \times \mathcal{L}_s \quad (1)$$

These weights were the best ones for a multi-objective optimization (Sener and Koltun 2018) over the following possible weight distributions: $\mathcal{L}_p 0.8 + \mathcal{L}_s 0.2$, $\mathcal{L}_p 0.6 + \mathcal{L}_s 0.4$ and $\mathcal{L}_p 0.5 + \mathcal{L}_s 0.5$.

Data

Our dataset $\mathcal{D}$ is made of around 6 thousand conversations among which we labeled $|\mathcal{D}_L| = 1,500$ conversations by two annotators during one month. This leaves $|\mathcal{D}_U| = 4,383$ conversations unlabeled, making the data to be used in a semi-supervised manner small relative to what is usual in the literature (Berthelot et al. 2019) but following the requirement for the unlabeled data to hold useful related information (van Engelen and Hoos 2020). However, it still represents approximatively three times the labeled set. Additional data statistics are shown in Table 1. The unavailability of public datasets containing two levels of annotations as well as labels on problem resolutions implies that we are restricted to working on our own confidential dataset, mandatory to apply our framework on a customer service use-case. In $\mathcal{D}_L$, the distribution of the different types of customer problems is the following: 249 aborted, 538 solved, 463 test_required, 198 out-of-scope and 52 no_solution.

Dataset	$\mathcal{D}_L$	$\mathcal{D}_U$
Language	French	French
Domain	Customer Service	Customer Service
Source	Real Data	Real Data
Nb. Conv	1,500	4,383
Speaker Role	Yes	Yes
Customer Satisfaction	Yes	No
Problem Status	Yes	No
Representation	Conversations	Messages

Table 1: Data statistics. Representation refers to how we represent the data for each subset: as conversations for the labeled set or as independent message for the unlabeled set used during the intermediate fine-tuning – step (3).

Experimental Protocol

We evaluate our approach by performing comparison of our model with several baselines, which are standard Fine-tuning, and two recent adaptation strategies from the literature: Domain-Adaptive Pre-training and Task-Adaptive Pre-training.

Fine-Tuning (FT). First, we apply standard Fine-Tuning (FT) on $\mathcal{D}_L$ to predict labels p and s separately using only an additional classification layer on top of the pooler layer of the pre-trained language model.

Domain-Adaptive Pre-training (DAPT). We apply DAPT (Konlea and Jannidisa 2020; Wu et al. 2021) with the standard Masked Language Modelling (MLM) objective, using the original BERT model masked/random tokens distribution (Devlin et al. 2019): 15% of the tokens are selected, among which 80% are replaced with the [MASK] token, 10% with a random token, and 10% not modified. DAPT can be summarized as follows: (1) pre-train the language model on a general domain dataset (2) continue training with the MLM objective on $\mathcal{D}_U$, not considering the labels a (3) fine-tune the language model on the target tasks to predict p and s, separately.

Task-Adaptive Pre-training (TAPT). We apply TAPT, which is also called intermediate fine-tuning (Howard and Ruder 2018; Gururangan et al. 2020) by simply fine-tuning our language model twice, adding this intermediate fine-tuning (with its own classification layer on top of the pooler layer) before fine-tuning on the target task. Hence, we follow Gururangan et al.'s implementation and use the following process: (1) pre-train the language model on a general domain dataset (2) fine-tune the model on $\mathcal{D}_U$ to predict the labels a (3) fine-tune the model for the target tasks to predict p and s, separately. TAPT can be seen as a trade-off between the limited FT and the excessive run time of DAPT.

Task-Adaptive Fine-tuning (TAFT). We then apply our approach, TAFT, with three different configurations. First, we apply TAFT without multitask learning. This means that contrarily to what is shown in Figure 1, where both the problem status p and the customer satisfaction s are predicted together, we predict them separately and hence execute the whole step (4) procedure twice. Secondly, we apply TAFT without the adaptive layer. This means that, skipping step (2), only the initial model is trained in step (3), and is not frozen, but fine-tuned in step (4). Lastly, we apply TAFT as originally described. During the multi-task fine-tuning of our TAFT model, we select the best model based on the weighted F1-score on validation set for the problem status classification task. This, along with our choice of weights in Equation 1, reinforces the 'primary' role of this task, and clearly directs our model towards favouring it.

Problem Status and Problematic Conversation. When predicting the status of the problem p, we ignore the 249 conversations cut short ("Aborted" label), in order to avoid unilateral or too short conversations, which cannot be exploited for predicting the status of the problem p nor the satisfaction s, and only consider the last four labels (described in Section Tasks and Labels Definition). In the following, we name this task **Status** or Problem Status. Besides, in order to allow us to check how efficiently we can identify problematic conversation, we propose a simplification of this task. We use a different mapping where conversations with the "Solved" and "Test_required" labels are considered *"unproblematic"*, while "out-of-scope" and "No solution" labels are merged together into a *"problematic"* label. This results in a binary Problematic Conversation identification task, shortened as **PC**. In the following, each configuration, whether it implies multi-task learning with the satisfaction s or not, is tested with **Status** and **PC**.

Speaker role identification. During the intermediate fine-tuning phase, we exploit all the unlabeled data without filtering on conversation length. Indeed, speaker role information is still available in any message and in this specific case, we do not consider the conversational context. Thus, each speaker role label y^a is only associated to its utterance u, which is hence taken in isolation and can be used to pre-train the adaptive layer in step (3) of the TAFT procedure or to fine-tune the language model in step (2) of the TAPT baseline whatever the length of the conversation (the DAPT baseline uses the same data as step (2), but without role labels).

Models and Hyperparameters. For our initial model and baselines, we use the distilled version of a RoBERTa (Liu et al. 2019) model for French, the CamemBERT model (Martin et al. 2020; Delestre and Amar 2022). The adaptive layer is a transformer encoder (Vaswani et al. 2017) with a feed forward dimension $d_{ff} = 2048$, a number of layers $N = 1$ and a number of attention heads of 12. The number of attention heads is chosen based on the self-attention block size d_k and the hidden dimension d_h of the initial language model: d_k/d_h. This adaptive layer is hence made of 16.5M parameters and takes as input the last hidden state h_{-1} from the RoBERTa model. In order to simplify its output for the classification tasks, we add a pooler layer to only consider the $[CLS]$ tokens representations and feed them to a linear layer with a hyperbolic tangent activation function $tanh(h_{-1}W + b)$. The adaptive layer is followed by a dropout layer with a droprate of 0.5 in order to prevent an over-fitting of the model on the different tasks. To first fine-tune the language model on the speaker role identification task, we tried different hyperparameters, especially in regards to different batch size (16, 32) and different learning rates ($2e-5$, $5e-6$, $5e-5$). We set the learning rate to $5e-5$ and the batch size to 16 while fine-tuning on 4 epochs over the unlabeled set. While the decay rate to the RoBERTa parameters is to 0.01, we lower the decay rate for the adaptive layer to 0.001 but consider a higher learning rate of $1e-4$. This choice stems from the DistilCamemBERT model containing 68.1M parameters while the adaptive layer only contains 16.5M parameters. We use the AdamW (Loshchilov and Hutter 2017) optimizer to train both the language model and the adaptive layer.

Unbalanced Strategies. We deal with a particularly small and highly unbalanced dataset. To address this limitation we implemented two different strategies : a pseudo episodic batch strategy, and probability draws to balance the whole dataset. The first strategy stems from the idea of episodic composition to represent data (Ravi and Larochelle 2017), often used in few-shot learning (Miller, Matsakis, and Viola 2000; Lake, Salakhutdinov, and Tenenbaum 2015). Usually, an episode is represented by a combination of three parameters: samples ($shots$) per class (way) to train from, and a set of samples per class to predict on ($queries$). We adjust this composition by only considering the number of samples to train from in order to force a balanced representation of the classes during training. We adopt the replacement strategy to draw samples, which means some samples from the

	Strategy	Satisfaction				PC				Status			
		P	R	F1	κ	P	R	F1	κ	P	R	F1	κ
Baselines	FT	49.09	47.20	43.07	0.19	52.77	52.83	52.68	0.05	45.02	*50.38*	*44.37*	*0.28*
	DAPT	46.35	50.40	46.34	0.13	**86.32**	**85.60**	**85.51**	**0.71**	47.44	32.40	29.17	0.18
	TAPT	47.22	52.33	46.34	0.13	73.82	71.23	70.69	0.43	33.85	44.34	38.09	0.25
Ours	TAFT noMTL	34.25	49.00	40.01	0.07	13.62	26.89	17.85	0.02	13.48	25.47	17.48	-0.01
	TAFT noAdapt	**67.74**	*59.38*	*57.47*	*0.41*	63.17	56.60	49.98	0.13	*49.01*	40.09	36.00	0.22
	TAFT	*63.21*	**62.80**	**62.92**	**0.44**	*74.39*	*73.58*	*73.67*	*0.47*	**53.78**	**52.36**	**51.88**	**0.36**

Table 2: Different strategies applied for three target tasks: customer's satisfaction prediction (Satisfaction), Problematic Conversation identification (PC), and Problem Status prediction (Status). Strategies include a fine tuning (FT) of DistilCamemBERT, a Domain-Adaptive Pretraining (DAPT), a Task-Adaptive Pretraining (TAPT), and our Task-Adaptive Fine-Tuning (TAPT) with or without Multi-Task Learning (noMTL), or without the adaptive layer (noAdapt). Each is evaluated using Precision, Recall, weighted F1-score, and Cohen's Kappa (κ) scores. Best scores are in bold, second best in italics.

lowest represented class will be seen more often; moreover, this strategy leads to additional hyper parameters, such as the number of episodes to draw and the number of shots, which tend to require dedicated algorithms (Snell, Swersky, and Zemel 2017; Guibon et al. 2021). The second strategy is to assign to each sample a draw probability from a multinomial distribution weighted by the label occurrence counts across the whole training set. This creates less overlap than the episodic composition, without required additional hyper parameters. Reported results refer to this strategy.

Results

Intended outcome. Ideally, we want our model to reach or outperform the Cohen's κ scores obtained during the manual annotation of the data, which only denoted moderate agreement at the conversation level (Landis and Koch 1977): the Cohen's κ score computes the agreement score between two annotators as $\kappa = \frac{p(a)-p(r)}{1-p(r)}$, taking into account the agreement probability between two annotators $p(a)$ and the random agreement probability $p(r)$. The annotation phase resulted in $\kappa = 0.76$ for the polarity of the visitor's satisfaction, $\kappa = 0.46$ for the problem status, and $\kappa = 0.55$ for the problematic conversation label. Table 2 shows the overall results we obtained in regards to precision, recall, weighted f1-score and Cohen's κ score: to compare models, we compute this κ score between the predicted values $\hat{Y}$ and the gold values Y as indicated by κ in Table 2.

Baselines Performance. Our baselines are a standard fine-tuning (FT), domain-adaptive pre-training (DAPT) and task-adaptive pre-training (TAPT); we evaluate them on customer satisfaction (Satisfaction), problematic conversation detection (PC) and customer problem status (Status). First, the fine-tuning shows very limited results on the Satisfaction and PC tasks; however, it achieves the second best on *Status* prediction with 44.37 in weighted F1-score. Secondly, DAPT does not outperform the fine-tuning on *Status* prediction but outperforms all the other approaches on detecting problematic conversations (PC). Since it achieves very poor results compared to the other baselines in the other tasks, we believe this is due to an interaction between the complexity of the tasks with the quantity of available data: DAPT is not able to extract sufficient information from $\mathcal{D}_U$ for the more complex labels. TAPT yields results equivalent to DAPT on customer satisfaction classification, which is the expected result – as one of the main purposes of TAPT was to reduce the computational cost of DAPT (Gururangan et al. 2020). Even though, κ scores are slightly improved by TAPT. Still, it is far behind DAPT on problematic conversation identification with a F1-score of 70.69. This seems to reinforce the intuition that TAPT is mainly useful for more complex tasks, by mean of its efficiency: the Masked Language Modelling process of DAPT induces a far higher training cost compared to TAPT which requires more data for simpler tasks. Still, both TAPT and DAPT come short when it comes to complex tasks for which data are scarce.

TAFT Performance. Our Task-Adaptive Fine-Tuning method (TAFT) outperforms both those baselines on Problem Status prediction by reaching 51.88% in f1-score, and on the Customer Satisfaction task with 62.92% in f1-score. This means our approach improves by 16 points on satisfaction and 7 points on status prediction. Status prediction being seemingly the most difficult task, our TAFT method reaches a satisfactory trade-off between the computational cost of DAPT and the efficient adaptation of TAPT. On the problematic conversation identification task, however, DAPT really outperforms all the other strategies and our TAFT only achieves the second best score, 12 points behind DAPT, but still 3 points above the TAPT. This strengthens the intuition that our TAFT is better suited for complex tasks when data is scarce. On *Status* prediction, our model still shows important flaws in predicting the positive satisfaction values along with the "No Solution" class. As we make sure data representation is approximately balanced, this may arise from the extreme definition of this specific label. Table 3 shows the exact same pattern on the *Status* prediction where we can see our model better predict clear opposite labels ("Solved", "No Solution") while struggling to predict more subtle ones ("To be tested", "Out-of-scope"). On the other hand, TAFT yields symmetric patterns between Precision and Recall scores while identifying Problematic Conversations (Table 3). On a side note, all the models achieve really high results ($\sim 93\%$ in f1-score)

	P	R	F1
Problem Status			
Solved	60.56	60.56	60.56
To be tested	58.33	30.43	40.00
Out-of-scope	36.67	46.81	41.12
No Solution	56.14	66.67	60.95
Weighted Avg	53.78	52.36	51.88
Standard Dev.	04.00	04.00	04.58
Problematic Conversation			
Not Problematic	67.59	77.66	72.28
Problematic	79.81	70.34	74.77
Weighted Avg	74.39	73.58	73.67
Standard Dev.	03.19	03.19	03.19

Table 3: TAFT scores per category on the Problem Status prediction and problematic conversation identification tasks, with their standard deviation over 5 test runs.

for the speaker role identification task during their respective pre-training phase.

Improvements from Multi-Task Learning In Table 2 we indicate the absence of target multi-task fine-tuning as "noMTL". It turns out the use of multi-task fine-tuning is an essential part of the performance of TAFT, as results for individual fine-tuning are very poor. This means the fine-tuning of the adaptive layer requires correlated tasks (Appendix A) to better transition from the adapting to the domain (step 1 in Figure 1) to adapting to the task (step 2 in Figure 1). The multi-task fine-tuning in TAFT has the added benefit to allow the model to have a better balanced performance: no problem status is completely ignored, contrary to DAPT and TAPT.

Improvements from the Adaptive layer. In Table 2 we indicate the absence of an adaptive layer with "noAdapt". Removing the adaptive layer but keeping the exact same process yields lower results on all the 3 tasks, and all metrics. This verifies the importance of the adaptive layer in our TAFT approach and shows that the adaptive layer enhances the adaptation process of the model – combining both domain and task adaptation at once.

TAFT Benefits. Overall, compared to the performance given by a random choice (0.25), the intended outcome ($k = 0.46$), and those three baselines, our model yields better results on the Problem Status prediction task. Even though, a lot of improvements are still required as the absolute score remains low. Compared to DAPT, our model is more efficient during the fine-tuning phase, as it trains on 16.5M parameters instead of 68.1M parameters. We think this is where our TAFT shines, by leveraging the relatively small unlabeled data in a more efficient manner. While most semi-supervised approaches can rely on a large quantity of unlabeled data, in our use case, where even unlabeled data is scarce, TAFT yields the best results compared to other inductive semi-supervised baselines.

Limitations

While our approach yields improvements, it has several limitations. The first is found in the nature of the problem status labels. These labels cannot simply be mapped to lexical features, representing specific topics, as there is a need to understand the evolution of a certain problem that the customer faced. Another possible labeling scheme would be to divide the Status into two sub-labels: the type of the problem, and its status. However, due to the cost of annotating, we chose to keep labels simple. The second limitation of our work is the use of a distilled RoBERTa for French. Even if we obtained few to no improvement from using the full CamemBERT model, other types of language models could be used in the TAFT approach. This has yet to be tested. Finally, the quantity of available data is obviously a big limitation in further experimenting with our model. It is a limitation that is inherent to our use-case of customer service – as are the difficult problem status labels. As often in customer service, the data is confidential and we are not able to share it.

Conclusion and Perspectives

In this paper, we presented a new inductive semi-supervised approach dedicated to datasets in which both the labeled and unlabeled subsets are relatively small. This approach is tested in a customer service use-case, on conversational data, leveraging speaker role information from the unlabeled set in order to better adapt the model to the target tasks, that are to predict customer satisfaction and identify the status of the customer problem – which we are the first to tackle.

Our Task-Adaptive Fine-Tuning approach achieves encouraging results compared to other recent inductive semi-supervised methods. Using a supplementary *adaptive layer*, it eases the adaptation of a pre-existing model to the task, in a two-steps process: first, both adapting the model to the new domain, and tying it to the new layer, and second, fine-tuning only the new layer to the target tasks. Additionally, multi-task training during that second step is an essential part of the model performance. In particular, our model outperforms baselines on the problem status task, showing that progress can be made on predicting complex labels even when data is very scarce. With this adaptive layer, our approach requires substantially less parameters to be trained during the final fine-tuning, and thus, while it combines aspects from task-adaptive and domain-adaptive fine-tuning, it yields a more efficient procedure than both. However, when we choose to simplify our target task to only identifying problematic conversations, our approach achieves competitive results but stays behind a pre-existing domain-adaptation strategy.

In the future, we plan to further improve problem status prediction by using argument mining mixed with sentiment analysis to identify the nature of the problem before its status, making it a two-step classification process. We will also consider a comparison with additional human performance.

Ethical Statement

In this work, it is important to consider this task as a help given to the operator. If we only base our decisions on automatic problem resolution status identification, then ethical

issues may arise based on how biased our model is. However, in this paper, we do not have access to clients identity and only consider two distinct speakers: client and operator.

Acknowledgements

This project has received funding from SNCF, the French National Research Agency's grant ANR-21-REVITALISE and the DSAIDIS chair at Télécom-Paris.

References

Ando, A.; Masumura, R.; Kamiyama, H.; Kobashikawa, S.; and Aono, Y. 2017. Hierarchical LSTMs with Joint Learning for Estimating Customer Satisfaction from Contact Center Calls. In *INTERSPEECH*, 1716–1720.

Auguste, J.; Charlet, D.; Damnati, G.; Béchet, F.; and Favre, B. 2019. Can we predict self-reported customer satisfaction from interactions? In *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 7385–7389. IEEE.

Bao, Y.; Ma, Q.; Wei, L.; Zhou, W.; and Hu, S. 2022. Speaker-Guided Encoder-Decoder Framework for Emotion Recognition in Conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Bothe, C.; Weber, C.; Magg, S.; and Wermter, S. 2018. A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Chang, T.-Y.; and Lu, C.-J. 2021. Rethinking Why Intermediate-Task Fine-Tuning Works. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 706–713. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Chen, D.; Chen, H.; Yang, Y.; Lin, A.; and Yu, Z. 2021. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3002–3017.

Chowdhury, S. A.; Stepanov, E. A.; Riccardi, G.; et al. 2016. Predicting User Satisfaction from Turn-Taking in Spoken Conversations. In *Interspeech*, 2910–2914.

Clavel, C.; Adda, G.; Cailliau, F.; Garnier-Rizet, M.; Cavet, A.; Chapuis, G.; Courcinous, S.; Danesi, C.; Daquo, A.-L.; Deldossi, M.; et al. 2013. Spontaneous speech and opinion detection: mining call-centre transcripts. *Language resources and evaluation*, 47(4): 1089–1125.

Delestre, C.; and Amar, A. 2022. DistilCamemBERT : une distillation du modèle français CamemBERT. In *Conférence sur l'Apprentissage automatique*. Vannes, France.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Gong, X.; Kong, X.; Zhang, Z.; Tan, L.; Zhang, Z.; and Shao, B. 2019. Customer service automatic answering system based on natural language processing. In *Proceedings of the 2019 International Symposium on Signal Processing Systems*.

Guibon, G.; Labeau, M.; Flamein, H.; Lefeuvre, L.; and Clavel, C. 2021. Few-Shot Emotion Recognition in Conversation with Sequential Prototypical Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. Online: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, 2122. NIH Public Access.

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328–339. Melbourne, Australia: Association for Computational Linguistics.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Jain, N. 2021. Customer Sentiment Analysis using Weak Supervision for Customer-Agent Chat. *arXiv preprint arXiv:2111.14282*.

Konlea, L.; and Jannidisa, F. 2020. Domain and Task Adaptive Pretraining for Language Models. *Proceedings http://ceur-ws. org ISSN*, 1613: 0073.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Landis, J. R.; and Koch, G. G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Lavalley, R.; Clavel, C.; Bellot, P.; and El-Beze, M. 2010. Combining text categorization and dialog modeling for speaker role identification on call center conversations. In

Eleventh Annual Conference of the International Speech Communication Association.

Liu, L.; Zhang, Z.; Zhao, H.; Zhou, X.; and Zhou, X. 2021a. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13406–13414.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021b. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Liu, X.; Shin, H.; and Burns, A. C. 2021. Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. *Journal of Business Research*, 125: 815–826.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 6818–6825.

Martin, L.; Muller, B.; Suárez, P. J. O.; Dupont, Y.; Romary, L.; De La Clergerie, É. V.; Seddah, D.; and Sagot, B. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219.

Maslowski, I.; Lagarde, D.; and Clavel, C. 2017. In-the-wild chatbot corpus: from opinion analysis to interaction problem detection. In *ICNLSSP 2017*, 115–120.

Miller, E. G.; Matsakis, N. E.; and Viola, P. A. 2000. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000*, volume 1. IEEE.

Park, K.; Cha, M.; and Rhim, E. 2018. Positivity bias in customer satisfaction ratings. In *Companion Proceedings of the The Web Conference 2018*, 631–638.

Park, K.; Kim, J.; Park, J.; Cha, M.; Nam, J.; Yoon, S.; and Rhim, E. 2015. Mining the minds of customers from online chat logs. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

Piris, Y.; and Gay, A.-C. 2021. Customer satisfaction and natural language processing. *Journal of Business Research*, 124: 264–271.

Prasad, A.; Zuluaga-Gomez, J.; Motlicek, P.; Ohneiser, O.; Helmke, H.; Sarfjoo, S.; and Nigmatulina, I. 2021. Grammar based identification of speaker role for improving atco and pilot asr. *arXiv preprint arXiv:2108.12175*.

Pruksachatkun, Y.; Phang, J.; Liu, H.; Htut, P. M.; Zhang, X.; Pang, R. Y.; Vania, C.; Kann, K.; and Bowman, S. R. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5231–5247. Online: Association for Computational Linguistics.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shang, G.; Tixier, A.; Vazirgiannis, M.; and Lorré, J.-P. 2020. Speaker-change Aware CRF for Dialogue Act Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, 450–464.

Shi, Y.; Lan, Z.; Liu, W.; and Bi, W. 2009. Extending semi-supervised learning methods for inductive transfer learning. In *2009 Ninth IEEE international conference on data mining*, 483–492. IEEE.

Shipitsin, V.; Bespalov, I.; and Dylov, D. V. 2022. GAFL: Global adaptive filtering layer for computer vision. *Computer Vision and Image Understanding*, 103519.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, 4077–4087.

van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wu, H.; Xu, K.; Song, L.; Jin, L.; Zhang, H.; and Song, L. 2021. Domain-Adaptive Pretraining Methods for Dialogue Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 665–669. Online: Association for Computational Linguistics.

Yang, L.; Shen, Y.; Mao, Y.; and Cai, L. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11595–11603.

Yuji, N.; Tetsuya, S.; and Yoshi, H. 2022. Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization. In *Proceedings of the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhou, H.-Y.; Oliver, A.; Wu, J.; and Zheng, Y. 2018. When semi-supervised learning meets transfer learning: Training strategies, models and datasets. *arXiv preprint arXiv:1812.05313*.