Performance of BERT on Persuasion for Good

Saumajit Saha, Kanika Kalra, Manasi Patwardhan, Shirish Karande
TCS Research Pune, India
{saha.saumajit, kalra.kanika, manasi.patwardhan, shirish.karande}@tcs.com

Abstract

We consider the task of automatically classifying the persuasion strategy employed by an utterance in a dialog. We base our work on the PERSUASION-FOR-GOOD dataset, which is composed of conversations between crowdworkers trying to convince each other to make donations to a charity. Currently, the best known performance on this dataset, for classification of persuader's strategy, is not derived by employing pretrained language models like BERT. We observe that a straightforward fine-tuning of BERT does not provide significant performance gain. Nevertheless, nonuniformly sampling to account for the class imbalance and a cost function enforcing a hierarchical probabilistic structure on the classes provides an absolute improvement of 10.79% F1 over the previously reported results. On the same dataset, we replicate the framework for classifying the persuadee's response.

1 Introduction

With the advancement of artificial intelligence, there has been a tremendous rise in its usage in the daily lives of people. Birth of conversational agents has made life a lot easier for organizations looking to achieve certain tasks. These agents, like as mentioned in (Luger and Sellen, 2016; Bickmore et al., 2016; Graesser et al., 2014), are goal-oriented, *i.e.*, they try to engage users in meaningful conversations and thereby aim to achieve their tasks. At times, they require different strategies of persuasion in order to mould people into their way of thinking, thereby changing their specific attitude or behaviour (Shi et al., 2020). Wang et al. (2019) proposed the foundation on building an automatic personalized persuasive dialogue system. They created the PERSUASION-FOR-GOOD dataset, which is composed of conversations between crowdworkers trying to convince each other to make donations to a charity. They annotated the utterances with persuasive strategy labels and proposed a baseline method for persuasive strategy classification.

Until recently, the dominant prototype in approaching any natural language processing tasks has been to focus on designing neural network architectures, using task specific data and word embeddings such as GloVe (Pennington et al., 2014). The NLP community is witnessing a paradigm shift towards pre-trained deep language representation model which achieves the SOTA in question answering, sentiment analysis and other NLP tasks. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) represents one of the latest developments in this field. It surpasses its predecessors, ELMo (Peters et al., 2018) and GPT (Radford et al.) by a significant margin on numerous NLP tasks. There is not much literature on exploring BERT for tasks related to persuasive dialogues.

In this work, we introduce a BERT-based approach to automatically classify the persuasion strategy employed by an utterance in a dialog. We also use the same approach to classify the type of the response of the persuadee's utterance. We base our work on the PERSUASION-FOR-GOOD dataset. Since the amount of annotated dialogues in this dataset are very less, we experiment to evaluate the efficacy of pretrained BERT in achieving better performance for the said task. The main contribution of this work is : 1. Creating a BERT-based hierarchical classification setup for classification of Persuader's strategy . 2. Creating a benchmark setup for the Persuadee's response classification 3. Additional analysis for the dataset introduced by (Wang et al., 2019).

The baseline performance for strategy classification on PERSUASION-FOR-GOOD is an F1 score of 59.6% and 74.8% accuracy (Wang et al., 2019). We observe that a straightforward fine-tuning of BERT does not provide significant performance gain: 60.60% F1 and 75.85% accuracy. Never-

Proceedings of the 18th International Conference on Natural Language Processing, pages 313–323
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

theless, non-uniformly sampling to account for the class imbalance does improve performance to 68.1% F1 and 77.69% accuracy. We further employed a cost function which enforces a hierarchical probabilistic structure on the classes, namely a utterance is Persuasive or Not, and if persuasive, the strategies further belong to coarser classes of Appeal or Inquiry. This step improves the performance to 70.39% F1 and 79.50% accuracy, which is an absolute improvement of 10.79 F1 over previously reported results.

The remainder of this paper is organized as: We review the related work in Section 2. In Section 3 we conduct data analysis. In Section 4 we describe our methodology. We analyze the results and present our observations in Section 5. Finally, we summarize the key conclusion in Section 6.

2 Related Work

2.1 Persuasive Conversation

Several previous works have looked at detecting persuasion in online forums and social networks. (Yu et al., 2019) performed fine-grained analysis of texts by detecting all fragments that contain propaganda techniques as well as their type in news articles. (Morio et al., 2019), (Tan et al., 2016) (Hidey and McKeown, 2018) worked on persuasion detection on online forum by modeling argument sequence in social media. (Yang et al., 2019) focuses on persuasive strategy detection in semi supervised fashion for sentences of posts on a crowd funding platform. Meanwhile, the number of papers which have attempted mining persuasive strategies in dialog conversations have been limited. (Keizer et al., 2017) evaluated persuasion as a strategy for negotiating dialog agent. However, we believe that the recent work by Wang et al. (2019) is a first attempt to explicitly collect corpus of persuasive conversations. They collect a persuasive conversation dataset(PERSUASION-FOR-GOOD) for charity donation with persuasive strategy annotations. Our work in this paper is based on this dataset. In recent years there has also been interest in generating persuasive utterances and slogans. (Munigala et al., 2018) generate persuasive captions for fashion items on an e-commerce website. (Li et al., 2019) generate dialogs based on the PERSUASION-FOR-GOOD dataset. Meanwhile, (Shi et al., 2020) have developed a retrieval based persuasive dialog agent with the same dataset. The scope of this work is however limited to strategy classification.

2.2 Hierarchical Classification

There are multiple ways the literature has exploited the hierarchical structure of the labels to improve the performance. (Kowsari et al., 2017) uses local models, viz. one for each node in the label hierarchy, where the lower level classifiers are stacked on top of the higher level. The inference is made using top-down strategy. There are flat approaches (Charuvaka and Rangwala, 2015; Xu and Geng, 2019), which employ one model per leaf node. They perform cost-sensitive classification by penalizing the mis-classification of negative examples as per their distance from the training class in the hierarchy. These approaches require multiple models for classification. Hence, using these approaches with the BERT based classification technique we have employed would be resource intensive.

There are techniques which take label embeddings into consideration. For example, (Rios and Kavuluru, 2018; Pal et al., 2020) approach of document classification, uses variants of Graph Neural Networks to embed the hierarchical information of the label space and employs information retrieval setting to match these label embeddings with the document vectors. In our method, Multilabel Classification with Probabilistic Structure (MLPS) described in section 4.2, we enforce the hierarchical probabilistic structure on the class predictions rather than the label embeddings.

Some local approaches (Gopal and Yang, 2013; Peng et al., 2018) employ regularization technique by constraining the parameters of the parent and child classifiers to be similar. Instead of using distinct set of parameters for the parent and the child, Banerjee et al. (2019) introduce inductive bias by initializing the parameters of the finer level classifiers with the parameters of the coarser level classifier and further fine-tuning them on the finer category classification. Our current baseline method for multil-label classification (ML), as described in section 4.2, does not take the advantage of the label hierarchies. In future we would like to extend this method to be on the lines of (Molino et al., 2018; Patidar et al., 2018), which model the label dependencies by using a sequential decoder to predict a branch of the labels in the hierarchy.

3 Dataset

Wang et al. (2019) designed an online task to col-

Binary Coarser Label	Ternary Coarser Label	Strategy Label
Persuasive	Persuasive Appeal	1) Emotional-appeal 2) Foot-in-the-door 3) Logical-appeal 4) Personal-story 5) Credibility-appeal 6) Donation-information 7) Self-modeling
	Persuasive Inquiry	8) Task-related-inquiry 9) Personal-related-inquiry 10) Source-related-inquiry
Non Persuasive	Non Persuasive	11) Non-strategy

Table 1: Persuader Strategy Label Hierarchy.

Coarser Label	Actual Label
Ask	1) Ask-org-info, 2) Ask-donation-procedure
Positive	3) Positive-to-inquiry, 4) Positive-reaction-to-donation, 5) Agree-donation
Negative	6) Disagree Donation, 7) Disagree Donation more, 8) Negative reaction to donation, 9) Negative to inquiry
Neutral	10) Neutral to inquiry, 11) Neutral reaction to donation
Greeting	12) Greeting
Other	13) Other, 14) Off-task, 15) Acknowledgement, 16) closing, 17) you-are-welcome, 18) thank
Task-related-inquiry	19) Provide donation amount, 20) confirm donation, 21) task-related-inquiry
Personal-related-inquiry	22) Ask persuader donation intention, 23) personal-related-inquiry

Table 2: Persuadee Response Label Hierarchy.

lect the persuasive dialog data. The main objective of the task was to persuade the other person to donate some amount for the charity *Save the Children*[1]. This task was performed by two participants, where one person who tries to persuade the other person for donation is termed as the persuader and the other person who donates is referred to as the persuadee. The persuader was instructed to use different types of strategies for persuading the persuadee. After the dialogue is over, both persuader and persuadee can choose to donate amount for charity. The chosen amount gets deducted from their task payment on Mturk.

This data collection process involved 1285 participants acting as either persuader or persuadee. After collecting the data, the authors annotated 300 dialogues out of 1017 with labels of persuasive strategy for each of the persuader's utterances in a dialogue. The annotated dataset consists of 10 persuasive strategies and one non-strategy class corresponding to persuader's utterances. The persuasive strategies listed in Table 1 were broadly categorized as persuasive appeal and persuasive inquiry. Each response of a persuadee was also annotated into one of the 23 different classes listed in Table 2.

The 300 dialogues annotated in Wang et al. (2019) are used to setup a persuasion strategy classification task. Each sample consists of the current persuader utterance and prior persuadee utterance which is considered as context. The sample has been labeled with one of the 11 strategy classes. We use these labels to associate coarser labels with each samples in accordance to Table 1. For e.g,

Context : That's so important. How do you raise donations?

Input : Do you currently donate to your charity?

Strategy label : Task-related inquiry

Coarser labels : Persuasive, Persuasive Inquiry

Wang et al. (2019) have provided a 5-fold data-split

[1] https://www.savethechildren.org/

such that the training set contains 3450 utterances and the validation set contains 863 utterances. The results presented in this paper, unless specified otherwise are based on these exact data splits.

We also address another task of Persuadee's response classification. We have created 5-fold splits from 300 annotated dialogues for persuadee response classification task in the manner similar as for Persuader's strategy classification. This contains 3877 samples in the training set and 970 samples in the validation set. Following is an example for such a task:

Context : Would $2.00 be too much to ask?

Input : No, I can do it.

Response type : Agree-donation

Coarser Label : Positive

3.1 Dataset Analysis

3.1.1 Dialog Independence

Wang et al. (2019) conducted a survey to categorize the personalities of the crowd workers, and did not notice a significant correlation with the choice of strategy. Nevertheless, if a human participates in several conversations then one could learn identity specific preferences for language usage and dialog strategy. We found out that 524 participants acted as only persuader, 584 acted as only persuadee and 177 participants acted as both. Figure 1 shows the count of conversation a participant participated in based on the acted role. This depicts that there are few participants which took part in more than one conversation. Hence, even though we believe that the modeling of identities can help personalize persuasion understanding, in this work, we do not take identities as input to our models, and train a single model which works only on utterances.

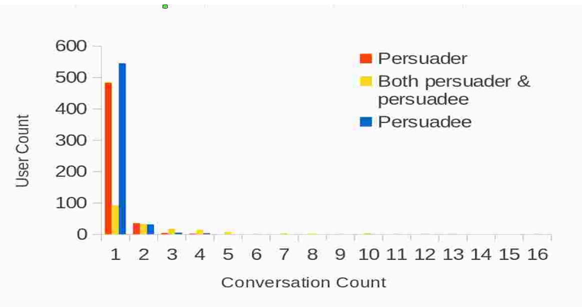

Figure 1: User role wise conversation participation count.

3.1.2 Interdependence of Responses

We also investigated for dependency among labels for both persuader and persuadee as well as current utterance and prior utterance labels individually for persuader and persuadee. We merged the persuadee class labels together to form 8 coarser class labels in the similar fashion as it was done for persuader strategies. Table 2 depicts the coarser label and the corresponding actual labels. We found out the sample mutual information by considering 8 persuadee class labels and 11 persuader strategy labels. The mutual information between persuader's current and prior utterance label is 0.1091 which indicates that persuader's current strategy is not influenced by prior strategy. The mutual information between persuadee's current and prior utterance label is 0.1222 which indicates that persuadee's current utterance response is not influenced by prior response. The mutual information between the persuader's current utterance label and persuadee's prior utterance label is 0.1452 which also indicates that persuader's strategy is not highly influenced by persuadee's response. These observations are reflected in our results section, where we did not observe significant advantages by including dialog history for classifying a particular utterance.

3.1.3 Truthfulness of Dialogues

We found that out of 643 persuadees participating in single conversation; only 355 donated and 288 did not donate. Further, out of the remaining 118 persuadees participating in more than one conversation; 41 donated in each of the conversation, 46 did not donate at all and rest donated in some of the conversations. Similar analysis was done for persuader's role as persuader can also agree to donate in conversation in order to persuade persuadee. We have found that out of 575 persuaders participating in single conversation; only 242 donated and 333 did not donate. Further, out of the remaining 126 persuaders participating in more than one conversation; only 39 donated in each of the conversation and 66 did not donate at all and rest donated in some of the conversations. In a dataset such as this, a particular conversation should ideally be considered persuasive in nature if and only if the persuadee donated amount for charity because of participating in the dialogue with persuader. Such causal analysis may prove particularly challenging as based on manual inspection of the data we noted that few workers did not make donations despite agreeing to do so in the conversation.

4 Methodology

4.1 BERT

The model architecture of BERT is a multilayer bidirectional Transformer encoder based on the original Transformer model (Vaswani et al., 2017). The input representation of the BERT can distinctly represent a pair of sentences as a sequence of tokens. For each token, its input representation is constructed by summing the wordpiece embedding (Wu et al., 2016), segment and the position embeddings. Segment embeddings help to distinguish one sentence from the other in the pair. A special classification [CLS] token is inserted in the beginning of the sequence. A separator [SEP] token is inserted at the end of each sentence in the pair. Finally, the final hidden state representation of the [CLS] token of each sequence can be used for the sentence classification tasks.

4.2 Multilabel Classification with Probabilistic Structure

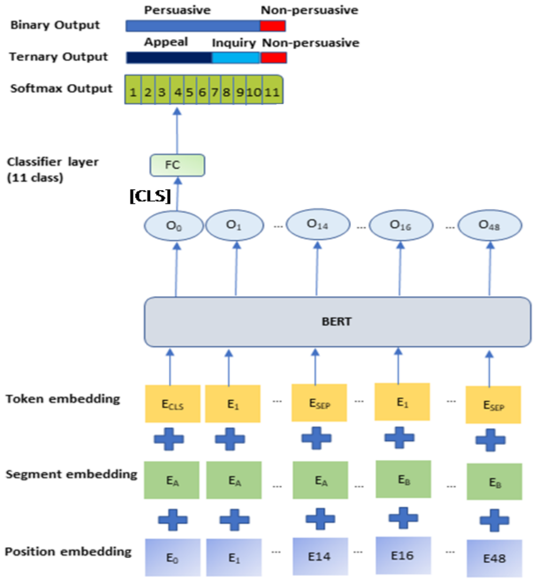

Figure 2: MLPS approach.

In persuader strategy classification task, input to BERT consists of the prior persuadee utterance as context and the current persuader's utterance. As shown in Figure 2 we use the final hidden state representation corresponding to the special [CLS] token as the aggregate representation of the input and pass it to a linear layer with *softmax* as its activation function. Finally, the posterior probability of each strategy is estimated by the softmax function P = **softmax**(WZ^T) where W is the weight matrix, $W \in \mathcal{R}^{d \times k}$ where k is the total number of strategies, d is the dimension of the [CLS] representation and Z is the representation of the final hidden state of the [CLS] token.

Softmax output is a k-dimensional vector, from which we choose the strategy corresponding to the highest value as our desired output. The first ten elements of this vector correspond to the persuasion strategies while the eleventh is the estimate for the probability that the utterance does not contain any persuasion. Furthermore, the first seven elements correspond to persuasive appeals. The posterior probabilities for the coarser labels can thus be estimated as:

$$P(Appeal) = \Sigma_{k=1:7} P_{\mathrm{k}} \tag{1}$$

$$P(Inquiry) = \Sigma_{k=8:10} P_{\mathrm{k}} \tag{2}$$

$$P(Persuasive) = \Sigma_{k=1:10} P_{\mathrm{k}} \tag{3}$$

During inference, at each label granularity, the label with largest estimate for the posterior probability is chosen. We highlight that one can train only on the 11 fine granular strategy classes, and yet conduct inference for all the coarser labels. We consider three label sets for training (a) $\lambda_{(11,\_,\_)}$ where only the 11 fine-granular strategy labels are used for training (b) $\lambda_{(11,\_,2)}$ where we additionally use the coarser binary label persuasive or not (c) and finally $\lambda_{(11,3,2)}$ where we further utilize the ternary labels (appeal, inquiry or non-persuasive) for training. The total loss is a weighted sum of the cross-entropy loss over labels at each granularity. Without loss of generality we enforce that the weights sum to one, and perform a grid search to identify the best performing weight combinations for above approaches (b) and (c). We restrict our search to the part of the grid where the weight for the binary classification is the highest. The intuition behind this is that if the network, once learns to correctly predict binary label as it is at coarser level, would further improve the multi-class fine granular prediction. In the remainder of the paper we refer to approaches (b) and (c) as multilabel classification with probabilistic structure (MLPS). A similar approach has been adopted for classification of Persuadee responses, and the output of softmax for 23 classes is binned together in accordance to Table 2. We consider an additional approach as shown in Figure 3 to create a baseline for multilabel classification, to understand the utility of specifically including hierarchy or label interdependency. In this baseline the output of the

[CLS] pin is given to label specific linear layers with dimension equal to number of classes associated with the label. In the remainder we refer to this approach as multilabel classification (ML).

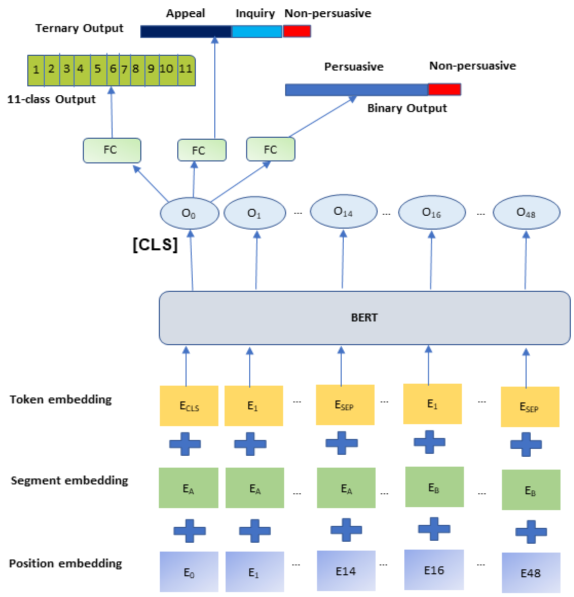

Figure 3: ML approach.

4.3 Turn Embedding

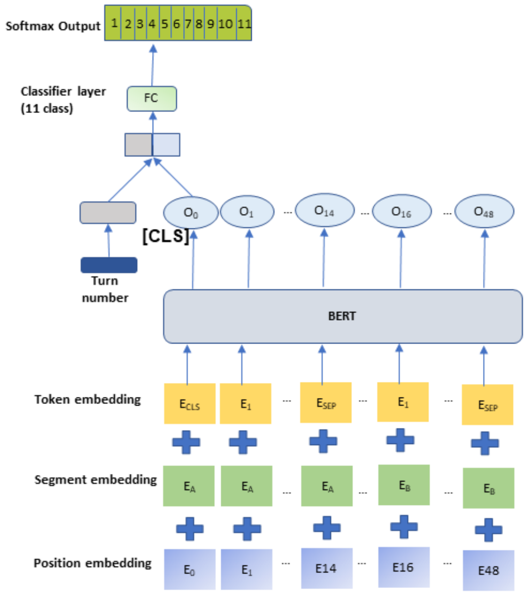

Figure 4: Architecture of Bert FT + Context + Turn

Wang et al. (2019) have shown that the distribution of employed strategies changes with the turn in a dialogue. We consider an ablation where we explore the utility of the turn side-information as an additional input to the persuader strategy classifier model. We consider a simple model where the 1-hot encoding for the turn is given to a hidden layer, which is concatenated to the [CLS] output of the final layer, before being fed to the linear layer before softmax. Turn embeddings did not prove beneficial when used in such a fashion, hence we did not consider them for multilabel classification. However, in future, it may be worth considering other approaches for embedding the turn information. Figure 4 refers to the approach which uses turn embedding.

5 Results and Analysis

5.1 Training Details

We use the pre-trained uncased BERT-base[2] model for fine-tuning. It consists of 12 Transformer blocks, its hidden layer size is 768, the number of self-attention heads present is 12, and the total number of parameters for the pretrained model is 110M. When fine-tuning, we keep the dropout rate to be 0.5, batch size to be 32 and the learning rate to be 2e -5. w_1, w_2 and w_3 are the weights of the loss functions, chosen in such a manner that they sum to 1. To tackle class imbalance, we have used Weighted Random Sampling (Efraimidis and Spirakis, 2008). We assign weights to the sampler such that each target label is assigned a weight equal to the reciprocal of the number of instances in the training set belonging to that target label. We have used Pytorch (Paszke et al., 2019) library while coding in python. Sequences of context and utterance having length greater than the maximum sequence length are truncated till 128. Shorter sequences are padded till the maximum sequence length. We use Early stopping in order to prevent over-fitting.

Our grid search revealed that: For persuader strategy classifier (a) for MLPS approach with training labels (11,_,2) best w_1 and w_2 are 0.4 and 0.6 respectively (b) for ML approach with training labels (11,_,2), w_1 and w_2 are 0.5 and 0.5 respectively (c) For MLPS approach with training labels (11,3,2), best result is found when w_1, w_2 and w_3 are 0.1, 0.3 and 0.6 respectively. (d) Similarly for ML approach with training labels (11,3,2), best result is obtained when w_1, w_2 and w_3 take the values 0.3, 0.3 and 0.4 respectively. For persuadee response classification (a) For MLPS approach, with training labels (23,8,_), best result is obtained when w_1 and w_2 are 0.1 and 0.9 respectively. (b) When w_1 and

[2] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

Persuader's Strategy Classifier							
Method	**# Training Labels**	**Acc-11**	**F1-11**	**Acc-3**	**F1-3**	**Acc-2**	**F1-2**
Hybrid RCNN (Wang et al., 2019)		74.8%	59.6%				
Bert FT	11,_,_	74.65%	66.36%	89.36%	88.13%	89.59%	88.69%
Bert FT + Context	11,_,_	77.69%	68.1%	89.48%	88.78%	90.46%	89.94%
Bert FT + Context + ML	11,_,2	77.82%	68.69%	89.39%	88.5%	89.88%	89.23%
Bert FT + Context + ML	11,3,2	78.41%	69.30%	89.81%	88.71%	91.11%	90.6%
Bert FT + Context + MLPS	11,_,2	78.61%	68.99%	90.11%	89.11%	91.13%	90.73%
Bert FT + Context + MLPS	11,3,2	**79.50%**	**70.39%**	**90.59%**	**90.07%**	**91.49%**	**91.08%**
Persuadee's Response Classifier							
Method	**# Training Labels**	**Acc-23**	**F1-23**	**Acc-8**	**F1-8**		
Bert FT + Context	23,_,_	57.66%	46.05%	68.00%	59.48%		
Bert FT + Context + ML	23,8,_	56.60%	47.08%	68.71%	62.44%		
Bert FT + Context + MLPS	23,8,_	**61.80%**	**54.18%**	**71.5%**	**66.63%**		

Table 3: Persuader Strategy Classification and Persuadee Response Classification Results (Acc-N: Accuracy for N classes, F1-N: F1 Score for N classes).

w_2 both take the value of 0.5, we get the best result with ML approach.

5.2 Ablation Study

Experiments	**Acc**	**F1**
Without sampler	75.85%	60.60%
With sampler	77.69%	68.1%
Only Utterance and no Context	74.65%	66.36%
Without Turn	76.03%	67.23%
With Turn	75.51%	67.64%
Persuasive/Non-persuasive classification with Utterance and History	88.63%	88.20%

Table 4: Results of ablation experiments for BERT baseline for single label.

We have conducted various ablation studies with the baseline Bert based classifier for persuader strategy classification. Table 4 shows the results of this study. We have trained model with and without weighted random sampler. As the dataset is highly imbalanced, we observed that non-uniform sampling improves F1-score significantly. Thus, all the experiments reported in Table 3 are with sampling using (Efraimidis and Spirakis, 2008). We have also seen the importance of context alongside the current utterance as input. There has been an improvement of 3.04% in accuracy and 1.74% in the F1-score. We have also incorporated turn information as shown in Figure 4 and observed no significant improvement. Finally, we trained the model only for binary classes and observed that the result improves when trained jointly with multilabel as reported in Table 3.

5.3 Multilabel Strategy classification

Table 3 presents the results for our experiments on multilabel classification. We observe that with non-uniform class sampling the baseline training for BERT on the 11 persuasion strategy classes, provides a performance of 66.36% F1. This is an improvement of 6.76% over the previously reported result. We also observe that even if label inter-dependencies are used only at inference time for the coarser labels, one can still get a performance comparable to directly training just for that label. The performance of both multiobjective approaches, ML and MLPS, was observed to be better than the BERT baseline approach. Thus including additional structure during training helps performance for all labels. We further observe that MLPS provides a performance better than ML demonstrating the utility of including probabilistic structure in the cost function. The best performance for classification on 11 persuasion strategies was observed to be 70.39% F1, with MLPS and labels (11,3,2). This is an improvement of 10.79% over the previously reported result, of 4.03% F1 over the BERT baseline and of 1.09% over the multilabel baseline.

As illustrated in Table 5, we calculated the class-wise F1-scores for persuader strategy classification with the baseline BERT-FT model, as well as ML and MLPS with $\lambda_{(11,3,2)}$ label set. We made the

Class Label	**BERT-FT**	**ML** (11,3,2)	**MLPS** (11,3,2)
Personal story	27.58%	43.25%	40.99%
Logical-appeal	46.47%	58.31%	58.89%
Task. inquiry	51.35%	54.99%	51.57%
Self-modeling	56.17%	74.01%	70.03%
Personal. inquiry	65.51%	67.31%	70.53%
Foot-in-the-door	68.75%	68.6%	67.44%
Emotional-appeal	74.13%	76.38%	73.89%
Donation info.	76.43%	78.47%	77.74%
Credibility-appeal	81.67%	83.61%	85.2%
Non-strategy	86.81%	85.76%	88.31%
Source. inquiry	87.09%	86.32%	90.68%

Table 5: F1 score for each of the class label in Persuader's strategy classification

following observations: (i) % change in F1-score over the baseline is positive for most of the classes for both MLPS and ML indicating both the methods provide a performance gain across classes (ii) the % gain in F1-score for MLPS is more evenly distributed across all the classes as compared to the gain for ML, and (iii) the classes with lower baseline F1-scores are roughly getting benefited more in MLPS setting than the ones with higher baseline F1-score. These observations for MLPS are consistent with those made in (Banerjee et al., 2019; Peng et al., 2018), which say that exploiting the hierarchical structure benefits all fine granular classes.

5.4 Response classification

Wang et al. (2019) have not provided any baseline for automatic classification of Persuadee's responses. We observe that the use of BERT with weighted sampling for class imbalance provides an accuracy of 57.66 % and F1 score of 46.05% over the 23 response classes. MLPS provides an improvement of 4.14 % in accuracy and 8.13 % in F1. MLPS also improves over the bases for the 8 coarser classes. The improvement in accuracy is 3.5 % and 7.15 % in F1. ML provides an improvement of 1.03% in F1 while the accuracy has slightly decreased for 23 classes. However for the coarser classes, accuracy has increased by 0.71% and F1 has improved by 2.96%. On the similar lines of persuader strategy classification, we calculated the class-wise F1-scores for persuadee response classification. We observed that the % gain in the F1-score is more evenly distributed (i.e. lower standard deviation) across all the labels for MLPS when compared to ML.

5.5 Comparative Analysis with Examples

There have been several instances where one approach outperforms another approach as shown in Table 5. This section provides some of the examples, in Persuader's strategy classification task, which highlight the performance of both the approaches for a given input.

Scenario where MLPS works better than ML

1. ***Context*** : $< Start >$
   Input : How much money do you spend daily on extras like a coffee or treat?
   Ground Truth : personal-related-inquiry
   MLPS : personal-related-inquiry
   ML : task-related-inquiry

2. ***Context*** : .60 still sounds good to me. Lets leave it at that.
   Input : I also want to assure you that Save the Children Makes huge impact on childrens lives internationally. They are extremely professional and your donation will go to a trustable fund.
   Ground Truth : credibility-appeal
   MLPS : credibility-appeal
   ML : logical-appeal

3. ***Context*** : I wish there was a long-term solution to these problems.
   Input : We all do, but for now, there are children in need and this organization does amazing work.
   Ground Truth : logical-appeal
   MLPS : logical-appeal
   ML : credibility-appeal

4. ***Context*** : Hi! Doing good. How are you?
   Input : I was wondering if I could talk to you about donating to Save the Children today?
   Ground Truth : source-related-inquiry
   MLPS : source-related-inquiry
   ML : task-related-inquiry

5. ***Context*** : I could donate 10 cents. I wish I could more but I am trying to pay my bills with what I make here.

Input : The children will thank you, do you think you can do a little more?
Ground Truth : non-strategy
MLPS : non-strategy
ML : task-related-inquiry

Scenario where ML works better than MLPS

1. ***Context*** : We sponsor a child in El Salvador, we have been going it a number of years.
Input : I donate money to save the children.
Ground Truth : personal-story
MLPS : self-modeling
ML : personal-story

2. ***Context*** : I spend about 5 dollars a day, you?
Input : Do you think that amount of money would make a difference in a needy child's life?
Ground Truth : task-related-inquiry
MLPS : personal-related-inquiry
ML : task-related-inquiry

3. ***Context*** : I will donate 10 cents of my 30 cent payment. You should type the same thing and if perhaps if you know anything about the charity, share it?
Input : This is a great charity and I will match you .10 cent payment.
Ground Truth : self-modeling
MLPS : non-strategy
ML : self-modeling

4. ***Context*** : I'll donate but not that much
Input : That's fine any amount helps.
Ground Truth : foot-in-the-door
MLPS : logical-appeal
ML : foot-in-the-door

6 Conclusion

We summarize the conclusions of our work as: Pretrained languages models like BERT may prove useful for natural language understanding of persuasion strategies even when data is scarce and imbalanced. Multilabel training which enforces a structure on the persuasion strategy class labels can help improve performance. A cost function based only on probabilistic structure was observed to provide the best performance. Probabilistic structure, even when used only during inference time, can provide competitive performance for coarser labels, which were not included in training. The performance gains due to MLPS were even more significant for classification of Persuadee's responses. MLPS offers more evenly distributed benefit for all the classes as compared to ML which can be more biased towards certain classes.

References

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300.

Timothy W Bickmore, Dina Utami, Robin Matsuyama, and Michael K Paasche-Orlow. 2016. Improving access to online health information with conversational agents: a randomized controlled experiment. *Journal of medical Internet research*, 18(1).

Anveshi Charuvaka and Huzefa Rangwala. 2015. Hiercost: Improving large scale hierarchical classification with cost sensitive learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 675–690. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pavlos Efraimidis and Paul Spirakis. 2008. *Weighted Random Sampling*, pages 1024–1027. Springer US, Boston, MA.

Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 257–265.

Arthur C Graesser, Haiying Li, and Carol Forsyth. 2014. Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, 23(5):374–380.

Christopher Thomas Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, Oliver Lemon, et al. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. ACL.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2019. End-to-end trainable non-collaborative dialog system. *arXiv preprint arXiv:1911.10742*.

Ewa Luger and Abigail Sellen. 2016. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297. ACM.

Piero Molino, Huaixiu Zheng, and Yi-Chia Wang. 2018. Cota: Improving the speed and accuracy of customer support through ranking and deep networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 586–595.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. 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 6275–6280.

Vitobha Munigala, Abhijit Mishra, Srikanth G Tamilselvam, Shreya Khare, Riddhiman Dasgupta, and Anush Sankaran. 2018. Persuaide! an adaptive persuasive text generation system for fashion domain. In *Companion Proceedings of the The Web Conference 2018*, pages 335–342.

Ankit Pal, Muru Selvakumar, and Malaikannan Sankarasubbu. 2020. Multi-label text classification using attention-based graph neural network. *arXiv preprint arXiv:2003.11644*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Mayur Patidar, Puneet Agarwal, Lovekesh Vig, and Gautam Shroff. 2018. Automatic conversational helpdesk solution using seq2seq and slot-filling models. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1967–1975.

Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 World Wide Web Conference*, pages 1063–1072.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access.

Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. *arXiv preprint arXiv:2001.04564*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system:

Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Changdong Xu and Xin Geng. 2019. Hierarchical classification based on label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5533–5540.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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)*, pages 3620–3630.

Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. 2019. Experiments in detecting persuasion techniques in the news. *arXiv preprint arXiv:1911.06815*.