Effective Forum Curation via Multi-task Learning

Faeze Brahman
University of California
Santa Cruz
fbrahman@ucsc.edu

Nikhil Varghese
University of California
Santa Cruz
nivarghe@ucsc.edu

Suma Bhat
University of Illinois at
Urbana-Champaign
spbhat2@illinois.edu

Snigdha Chaturvedi
University of North Carolina at
Chapel Hill
snigdha@cs.unc.edu

ABSTRACT

Despite several advantages of online education, lack of effective student-instructor interaction, especially when students need timely help, poses significant pedagogical challenges. Motivated by this, we address the problems of automatically identifying posts that express confusion or urgency from Massive Open Online Course (MOOC) forums. To this end, we first investigate the extent to which the tasks of confusion detection and urgency detection are correlated so as to explore the possibility of utilizing a multitasking set-up. We then propose two LSTM-based multitask learning frameworks to leverage shared information and transfer knowledge across these related tasks. Our experiments demonstrate that the approaches improve over single-task models. Our best-performing model is especially useful in identifying posts that express both confusion and urgency, which can be of particular relevance for forum curation.

1. INTRODUCTION

Massive online courses have changed the academic landscape of today, offering convenient alternatives to learners at significantly reduced costs, compared to traditional educational institutions. With more than six million students taking at least one online course as part of their degree program [16], online education has already become one of the most popular higher education supplements.

Despite several advantages associated with online education, such as diversity of programs, lower cost, and more flexible learning environment, factors such as lack of personalization and low instructor-student ratio pose significant challenges to this learning environment. For the most part, discussion forums continue to be the sole platform for student interaction with others (students and instructors), where learners share their ideas, opinions, or even express their concerns and questions about the course material. Unfortunately, in a typical online class, these forums can quickly get difficult to manage with few instructors and several learners getting involved and posting their concerns. This situation can hamper the instructors' ability to gauge students' comprehension of course materials and address students' concerns in a timely manner, ultimately reducing learning effectiveness for students.

One way of bringing about the much needed immediacy is by way of automatic curation of the forums, where posts related to confusion about the course material, or those that raise urgent issues are automatically identified. For instance, identifying posts that express confusion (*Confusion Detection*) could help instructors in adapting their teaching strategies during the course by employing more examples, altering the course syllabus or slowing down the pace of instruction. Likewise, automatically identifying urgent posts, i.e. posts which need an immediate response (*Urgency Detection*) and resolving them in a timely manner is important for keeping students engaged. The two types of posts are related but different in the sense that posts that express confusion seek help about the content of the course material while posts that express urgency also seek help but not necessarily directly about the course content. Nevertheless, the ultimate goal of both types of posts is to seek help from others and so there is promise in designing methods that can learn them simultaneously in a multi-tasking set-up.

While previous works have focused on addressing a single forum curation task [1, 20, 21, 22], other studies [24, 25] have also shown that learning features that help address one task may be gainfully used for other tasks–an aspect central to a multi-task learning framework. Another reason for exploring multi-task learning in this domain is the limited availability of labeled data. The use of supervised machine learning approaches requires labeled data annotated by experts, which can be time-consuming, costly, and difficult to obtain in this domain. Unlike single-task frameworks which often suffer from insufficient annotated data, the proposed multi-task framework can share information between related tasks leveraging beneficial information, thus avoiding the need to have large amount of labeled data for individual tasks. However, this comes at the cost of increased model-parameters, which can instead hurt the model. Also, if the jointly learned tasks are weakly correlated, it might be more fruitful to focus on one task at a time since multi-tasking might introduce more noise than useful signals. Despite these issues,

Faeze Brahman, Nikhil Varghese, Suma Bhat and Snigdha Chaturvedi "Effective Forum Curation via Multi-task Learning" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 356 - 363

the potential gains via an implicit increase in the sample size for training our model by making it learn related tasks has the promise of averaging the noise of each task and thus improving generalization.

In this paper, we propose two multi-task learning architectures, namely Shared-BiLSTM and Specific-Shared Multi-Task, based on Long Short Term Memory (LSTM) networks. Our goal is to use these architectures for forum curation by jointly learning the tasks of Confusion detection and Urgency detection. To investigate the potential promise of our approaches, in light of the concerns mentioned above, we design experiments to answer the following research questions:

RQ1: To what extent are different tasks in this domain correlated?
RQ2: What is an effective multi-task learning architecture for this problem?
RQ3: Can the proposed multi-task learning model leverage the shared signals between the correlated tasks?
RQ4: How does adding more tasks affect the model's performance in the primary tasks?
RQ5: Does an already trained multi-task model help in improving recall in a specific subset of data that could be of particular interest to the instructors? [1]

Our experiments show that automatic forum curation benefits from sharing signals between Confusion and Urgency detection, and our proposed multi-task learning architecture improves on the individual tasks by learning shared and mutually beneficial features between the tasks. We summarize our contributions as follows:

- We empirically explore the extent to which confusion and urgency detection are correlated using representative MOOC forum posts.
- We propose two multi-task learning architectures that share information between related tasks.
- Using representative forum posts, we empirically demonstrate that multi-task models improve over single-task models. Our proposed model is especially useful in detecting posts that express both confusion and urgency, which can be particularly relevant for forum curation.

2. RELATED WORK

As MOOCs have attracted millions of users worldwide, analyzing big data from online courses have become an indispensable means towards understanding students' learning patterns. In this regard, previous research has proposed models to predict dropout or success [7, 13, 14, 18], to measure the impact of social factors in attrition prediction [15], and to automatically curate discussion forums [2, 3, 4]. For example, Ramesh et al. [14] proposed a latent representation model which could be used to abstract student engagement types and to predict dropouts. Wang et al. [19] adopted a content analysis approach to investigate the relationship between students' cognitive behavior in MOOCs forums and their learning gains. Chaturvedi et al. [4] proposed chain-based models that incorporate meta-data along with course information and content of the posts to identify the posts

[1]These are instances where posts are labeled as both Urgent and Confusion.

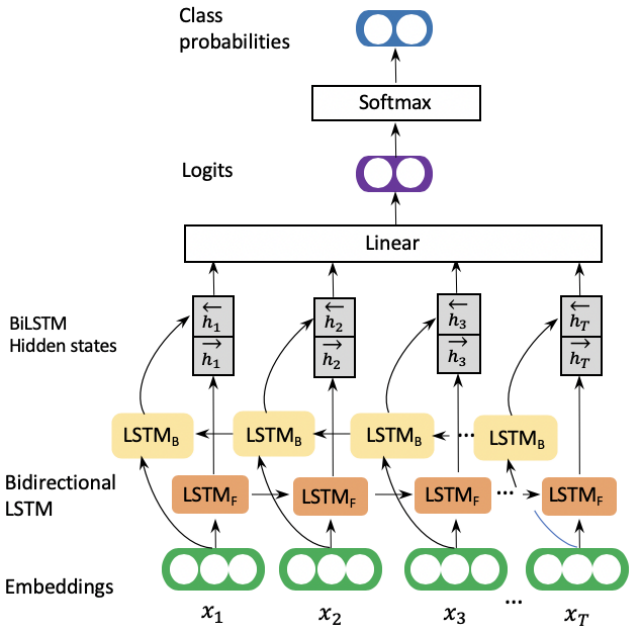

Figure 1: Single-Task Bidirectional LSTM Model.

that require instructor's attention. Chandrasekaran et al. [3] demonstrated the importance of prior knowledge about forum types in enhancing the predictive performance on the instructor's intervention task. Chandrasekaran et al. [2] proposed a supervised classifier which makes use of an automatic discourse parser for robust instructor intervention prediction.

Previous work has also focused on using behavioral and community-related cues to provide an insight into students' intentions, performances, and comprehension levels [21, 21]. Zeng et al. [22] and Agrawal et al. [1] investigated linguistic features along with structural features (e.g., the number of times a post has been read or the number of up-votes) to detect confusion. As identified by previous works [1, 22], one of the primary challenges in this area, is the lack of labeled instances and previous methods have explored the use of domain adaptation for addressing this challenge [23].

To address the problem of labeled data scarcity and leverage the relatedness between tasks, we propose to use multi-task learning which has been proven to perform well in many NLP tasks that include sequence labeling [5], text classification [10], machine translation [6]. For example, Liu et al. [9] proposed different architectures to control the information flow between shared or specific embedding and LSTM layers for text classification. However, multi-task learning has not been effectively explored for the online education domain. In this paper, we propose two multi-task frameworks to jointly learn related tasks (*confusion* and *urgency* detection) from the shared signals.

3. METHODOLOGY

We first define our task in Section 3.1 and in the following sections, we describe the Single-Task (ST), Shared-BiLSTM, and Specific-Shared Multi-Task (SSMT) models.

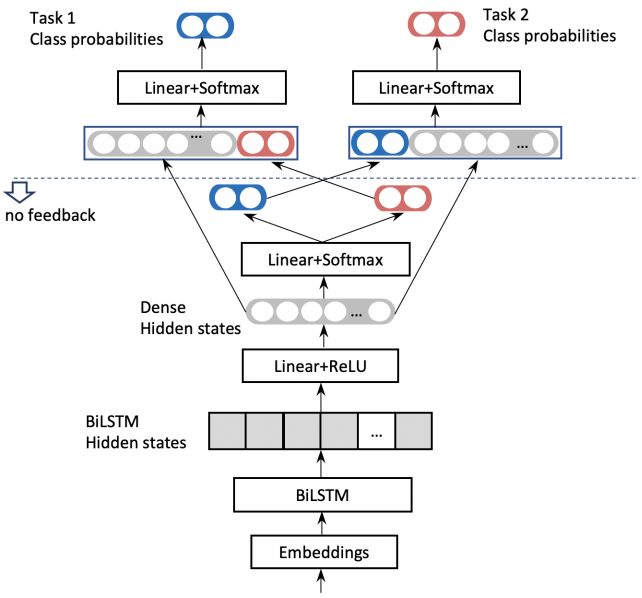

Figure 2: Shared-BiLSTM Model Architecture.

3.1 Problem Formulation

Our training dataset is $D = \{(X^i, Y^i)\}_{i=0}^{N}$, where X^i represents the i^{th} instance, and $Y^i = \{y_1^i, y_2^i, ..., y_M^i\}$ denotes a set of M labels for the instance, one corresponding to each task[2]. We assume that each task is a binary classification problem ($y_j^i \in \{0, 1\}$), but the proposed method can also work for multi-class classification tasks. In the following sections, we describe our different architectures.

3.2 Single-Task (ST)

We first create single-task models with identical architectures, to address the individual tasks of detecting confusion and urgency separately. The architecture is depicted in Figure 1. Given a forum post instance as a sequence of tokens $X^i = \{x_1, x_2, ..., x_T\}$, and the class label Y^i, we first use an embedding layer to get the vector representation of each token x_t, followed by a BiLSTM layer and a linear layer with softmax activation to obtain class probabilities. The model is trained to minimize the cross-entropy loss for each task:

$$L = -\sum_{i=1}^{N} y^i \log(\hat{y}^i) \tag{1}$$

Where y and $\hat{y}$ are the ground-truth and predicted labels (for a particular task) respectively.

3.3 Shared-BiLSTM

We now describe our first multi-task model that uses a shared BiLSTM encoder between different tasks to capture related information. The shared encoder has its architecture nearly identical to the single-task model except that it has an extra linear layer with ReLU activation between the BiLSTM and the Linear (with softmax) layers. Figure 2 shows the model architecture for two tasks, however; it can be easily extended for M tasks. Note that in this (and the next) figure certain layers are collapsed into one single layer for simplicity. For instance, we depict Linear and Softmax as Linear+Softmax in Figure 2 and 3. We experimented with two main variations of this architecture: (1) Without feedback, and (2) With feedback. The first variation, without feedback, is the part of the model shown below the dotted line in Figure 2. The second variation, with feedback, has the class probabilities of each task concatenated with the dense hidden states (the entire Figure 2). Given the training pairs of a post sequence $X^i = \{x_1, x_2, ..., x_T\}$, and the class label Y^i, the parameters of the model are updated to minimize total cross-entropy loss for the M tasks:

$$L_{total} = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_j^i \log(\hat{y}_j^i) \tag{2}$$

[2]In our case, we chose $M = 2$, where each label indicates if a post pertains to *confusion* and *urgency*.

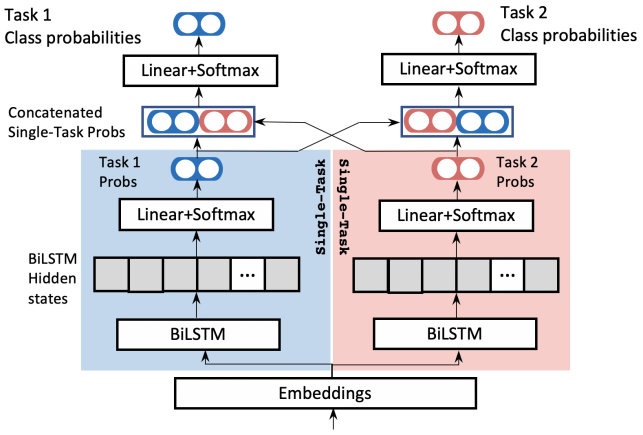

Figure 3: Specific-Shared Multi-Task (SSMT) Model Architecture

3.4 Specific-Shared Multi-Task (SSMT)

We now describe our second multi-task model, Specific-Shared Multi-Task SSMT, that unlike the Shared BiLSTM model, first models task-specific characteristics and then shares information between the tasks. This model has *task-specific* components, with architectures identical to that of single-task models, to learn task-specific features (shown in highlighted parts of Figure 3)[3]. Thereafter, the model shares information across tasks by concatenating the predictions of the task-specific components followed by a fully connected layer (with softmax activation) to make predictions for the various tasks. Given the training pairs of post sequence $X^i = \{x_1, x_2, ..., x_T\}$, and corresponding class labels Y^i, we first trained two separate single-task models, and used them to initialize the *task-specific* components of the multi-task network. We then trained the entire network to minimize the total cross-entropy loss defined in Equation 2. Note that during training, *task-specific* BiLSTM parameters get updated along with other model parameters.

4. EVALUATION

In this section, we evaluate the utility of the proposed multi-task models to address our primary tasks: Confusion and

[3]Like before, Figure 3 shows the architecture for two tasks, but can be easily extended for more tasks.

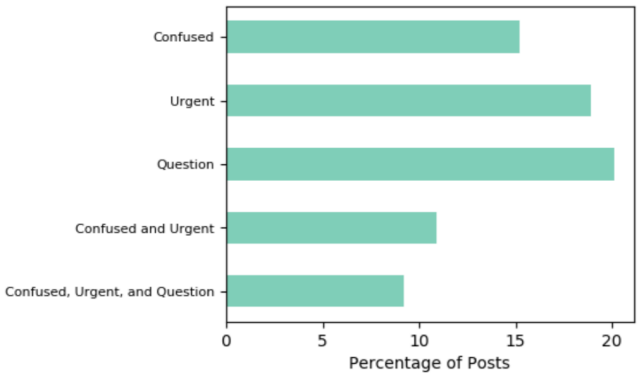

Figure 4: Label Distribution for the Stanford MOOC Posts Corpus

Urgency detection. Following previous works, we measure performance using Precision, Recall, and F1 scores of the positive class (*confusion or urgency*). This is because from the perspective of forum curation and helping students, positive class is more important than the negative class.

Dataset. We perform our experiments on the Stanford MOOC Posts Corpus [1]. The dataset contains 29,604 anonymized forum posts from 11 Stanford University public online classes spanning three broad domains: Humanities/Sciences, Medicine and Education. While this dataset has several labels, we primarily focus on two labels: Confusion and Urgency, labeled on a scale of $1-7$. The confusion rating is based on the extent to which the post expresses confusion, such as an inability to understand some concept that is taught in the class. Similarly, the urgency rating is based on how urgent it is that the instructors respond to the post. Although these labels are on a scale of $1-7$, following previous work [1], we convert these labels to binary values – posts with a score greater than 4 are categorized as *Confusion* (or *Urgency*), and those with a score equal or less than 4 as *Not Confusion* (or *Not Urgency*). Additionally, in some of our experiments, we use an additional label – *Question*, indicating whether the post was a question or not. Figure 4 shows the dataset's label distribution. We can see that only 15.19% of posts are labeled as *Confusion*, which shows a severe class imbalance in this dataset. We use an $80-10-10$ split for training, validation, and test data.

Training Details. For all our models, we initialized the embedding matrix with pre-trained 100-dimensional GloVe vectors [12]. We use a one-layer BiLSTM network with 80 hidden units. We experimented with using more layers and hidden units. However, that led to over-fitting possibly because of the relatively smaller size of the dataset. We applied dropout [17] of rate 0.2 between the BiLSTM hidden layers and the output layers for regularization, and did not fine-tune the word embeddings during training to avoid overfitting. Finally, we optimized using the Adam optimizer [8], with a learning rate of 0.001.

Correlation Analysis. We performed inter-label correlation analysis prior to our main experiments. First, we visualize the relationship between Confusion and Urgency (considering the original $(1-7)$ Likert scale) in the boxen plot shown in Figure 5. We can see that there can be disagreement between confusion and urgency labels especially around the threshold rating of 4. For example, there are several posts with confusion rating of 4.5 which would be labeled as *Confusion* but not *Urgency* (because their urgency ratings are less that 4). However, we observe a relatively high correlation between the two tasks for the most part.

Next, we also analyze the Spearman correlation between *confusion* and *urgency* (Table 1). We consider both original as well as the binary labels based on the threshold described earlier. We observe a moderate correlation between Confusion and Urgency (0.570). We also report correlations of these labels with respect to Question to explore whether it can be additionally used in the multi-task setup to improve the performance of Confusion and Urgency detection (the two primary tasks we are interested in). We also find that using binary labels increases the inter-label correlation for all cases. Note that inter-label correlation suggests but does not guarantee or quantify improvement in predictive performance with multi-task learning. Hence, in the following section, we design a new experiment where we consider three single-task models (confusion, urgency, and question) and explore the utility of each to predict Confusion and Urgency (RQ1). We then conduct other experiments to further investigate the utility of multi-tasking for these problems.

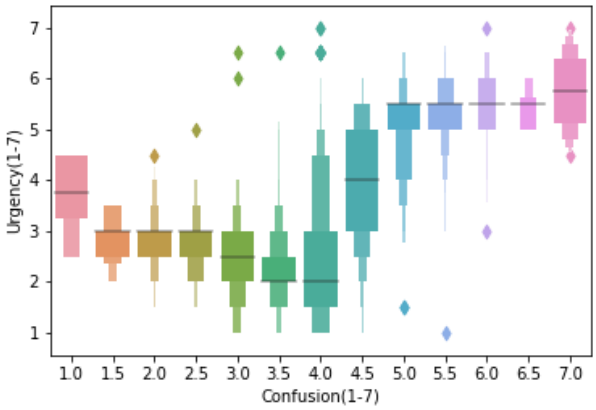

Figure 5: Inter-label correlation distribution between ordinal Confusion and Urgency label; the Spearman correlation value is 0.481**.**

4.1 Experimental Results

In our experiments, we implemented a single-task architecture mentioned in Section 3.2 to create models for each of the three tasks by training them on labeled data from the respective tasks: Single-Task Confusion detection (ST-C), Single-Task Urgency detection (ST-U), and Single-Task Question detection (ST-Q). These form our baselines. We follow a similar naming convention for the Shared-Specific Multi-Task model. For example, we refer to the Shared-Specific Multi-Task model to predict confusion and urgency together as SSMT-CU.

As a preliminary experiment, we compare the performances of our neural Single-Task models with Logistic Regression (LR) using Bag-of-Words and tf-idf features. Comparing the results in Table 2 with those in Rows 1 and 4 of Table 3, we

Labels	Confusion(1-7)	Confusion(1/0)	Urgency(1-7)	Urgency(1/0)	Question(1/0)
Confusion(1-7)	1.0	0.722	0.481	0.545	0.510
Confusion(1/0)	0.722	1.0	0.603	0.570	0.567
Urgency(1-7)	0.481	0.603	1.0	0.852	0.671
Urgency(1/0)	0.545	0.570	0.852	1.0	0.690
Question(1/0)	0.510	0.567	0.671	0.690	1.0

Table 1: Spearman correlation between all labels

Model	Task predicted	F1	Precision	Recall
LR-C+BOW	Confusion	0.45	0.56	0.38
LR-C+tf-idf	Confusion	0.38	0.68	0.27
LR-U+BOW	Urgency	0.61	0.67	0.57
LR-U+tf-idf	Urgency	0.59	0.76	0.48

Table 2: Performance evaluation of single-task models with Logistic Regression as baseline

Model	Task predicted	F1	Precision	Recall
ST-C	Confusion	0.50	0.68	0.40
ST-U	Confusion	**0.47**	0.46	0.48
ST-Q	Confusion	0.32	0.39	0.27
ST-U	Urgency	0.67	0.72	0.62
ST-C	Urgency	**0.44**	0.67	0.33
ST-Q	Urgency	0.44	0.60	0.47

Table 3: Performance evaluation of single-task models when used to predict *Confusion* or *Urgency*

can see that ST-U and ST-C outperform Logistic Regression based models on both the tasks. So, henceforth we use our neural models for all single task experiments.

RQ1: To what extent are different tasks in this domain correlated?
The goal of our first experiment is to find out if the tasks are correlated enough that model trained on one task can yield reasonable predictive performance on the other task. This would indicate if multi-tasking can help for jointly learning these tasks. For this purpose, we first evaluate ST-C, ST-U, and ST-Q on the task of confusion detection. Even though ST-U and ST-Q were not trained on this label (confusion), we posit that since the tasks of urgency and question detection are correlated with that of confusion detection, these models could have learned signals commonly shared with the confusion detection task. We perform a similar experiment to find correlations with urgency detection. All results are reported in Table 3.

The experiment indicates that the strongest correlation exists between the primary tasks: Confusion and Urgency detection. When used to predict the confusion label, ST-U obtains an F1 score of 0.47, which is only slightly lower than that obtained by ST-C (0.50). Similarly, ST-C performs relatively well in the urgency detection task suggesting that ST-U and ST-C have learned mutually beneficial signals, and can be used in a multi-task setup.

On the other hand, according to row 3 of Table 3, ST-Q has not learn enough mutually beneficial signals for the confusion detection task , suggesting that confusion and urgency are more useful for each other than question.

RQ2: What is an effective multi-task learning architecture for this problem?
We experimented with various versions of the two multitasking architectures proposed in Section 3. Here, we summarize these architectures and their performances.

For the Shared-BiLSTM model, we consider a variation without feedback (see Section 3.3) and three others with feedback. For the variations with feedback, we experimented with (1) Initializing the entire network randomly, (2) Pretraining and then freezing the shared encoder, and (3) Pretraining the shared encoder but further tuning the entire model to minimize total loss. Together these make up a total of 4 variations of the Shared-BiLSTM model. The performances are reported in the top half of Table 4. We can see that the variations which performed the best are the one that includes feedback with random initialization and the one with feedback, pre-training and freezing.

We also experimented three variations of SSMT: (1) Adding an extra Linear layer with ReLU activation between BiLSTM and final Linear Layers, (2) Including single-task losses in Equation 2 when fine-tuning the entire network, and (3) The model described in Section 3.4 without any changes. The results are summarised in lower half of Table 4. We can see that the model without any changes (as described in Section 3.4) outperforms its other two variations as well as all variations of the Shared-BiLSTM architecture. For the rest of our experiments we use SSMT as our final multitask setup and we discuss its performance in the rest of the research questions.

RQ3: Can the Specific-Shared Multi-Task model leverage the shared signals between the correlated tasks?
We evaluate our Specific-Shared Multi-Task model for predicting Confusion and Urgency (SSMT-CU). Table 5 shows that SSMT-CU outperforms both ST-C and ST-U on the two primary tasks. Comparing Rows 1 and 3, there is an increase in F1 score for the confusion detection task from 0.50 to 0.56. Comparing Rows 4 and 6 shows that we also obtain a boost in F1 score of the urgency detection task from 0.67 to 0.69. These results are statistically significant ($p < 0.001$) [11], and indicate that SSMT-CU has learned the shared signals between the two tasks. Also, we see that urgency has helped to identify confusion more than vice versa. This can also be observed in Table 3: the drop in performance when using ST-U instead of ST-C for confusion detection was much smaller than the drop resulting from using ST-C instead of ST-U for urgency detection. This also hints that urgency signals are more useful for confusion detection

	Confusion			Urgency		
	F1	Precision	Recall	F1	Precision	Recall
Shared-BiLSTM (w/o fb)	0.50	0.64	0.41	0.66	0.65	0.68
Shared-BiLSTM (+fbrandom-initialization)	**0.53**	0.67	0.44	0.63	0.70	0.57
Shared-BiLSTM (+fb+pre-training+freeze)	**0.53**	0.67	0.43	**0.67**	0.69	0.66
Shared-BiLSTM (+fb+pre-training+tune)	0.48	0.72	0.35	0.63	0.72	0.57
Specific-Shared Multi-Task (+dense)	0.52	0.62	0.44	0.68	0.66	0.70
Specific-Shared Multi-Task (+st-losses)	0.52	0.66	0.42	0.68	0.70	0.67
Specific-Shared Multi-Task	**0.56**	0.66	0.49	**0.69**	0.70	0.69

Table 4: Results of different variations of our two multi-task architectures. We indicate feedback with "fb", and single-task with "st". Bold fonts denote best performances among top and bottom halves of the table.

Model	Task predicted	F1	Precision	Recall
ST-C	Confusion	0.50	0.68	0.40
SSMT-CUQ	Confusion	0.52	0.71	0.41
SSMT-CU	Confusion	**0.56**	0.66	0.49
ST-U	Urgency	0.67	0.72	0.62
SSMT-CUQ	Urgency	**0.69**	0.71	0.67
SSMT-CU	Urgency	**0.69**	0.70	0.69

Table 5: Performance evaluation of single-task and multi-task models; MT-CU and SSMT-CUQ outperform ST-C and ST-U in the primary tasks.

than confusion signals for urgency detection.

RQ4: How does adding more tasks affect the model's performance in the primary tasks?
To investigate if adding the task of Question Detection can supplement the primary tasks, we introduce the SSMT-CUQ model and compare it with the existing models. Comparing Row 1 with 2, and 4 with 5 in Table 5, we find that SSMT-CUQ has a better F1 score than both ST-C and ST-U. This shows that adding an extra task still yields better performance than single-task models for the primary tasks.

To evaluate whether it further enhanced the SSMT-CU model, we compare Rows 2 with 3 and 5 with 6. SSMT-CU obtains a higher F1 score (0.56) than SSMT-CUQ (0.52) for the confusion detection task. We attribute the drop in performance of SSMT-CUQ for the confusion task to the relatively weaker correlation between the question detection and confusion detection tasks (also observed in our earlier experiment when comparing Rows 1 and 3 of Table 3). The introduction of question detection task might have introduced more noise and weakened the shared signals of confusion and urgency.

On the other hand, SSMT-CUQ and SSMT-CU have identical F1 scores (0.69) on the urgency detection task (Rows 5 and 6 of Table 5). Despite question detection being as useful for urgency detection as confusion detection (shown in Table 3), SSMT-CUQ did not improve over SSMT-CU because it might have received similar signals from both confusion detection and question detection tasks.

RQ5: Does an already trained multi-task model help improving recall in an specific subset of data that could be of particular interest to the instructors?
We now turn our attention to a specific subset of our dataset – posts labeled as both urgent as well as expressing confusion – for their potential to impact learner satisfaction [4]. In this experiment, the models are not trained on this subset. Instead, we analyze the performance of the (already trained) multi-task model on this subset. Since all the posts in this subset are labeled as *Confusion* and *Urgency*, any model will have a precision of 1 leading to a less informative F1 score. So, in this experiment, we focus on Recall values. Table 6 shows that the Specific-Shared Multi-Task model significantly outperforms the single-task models in the subset for both confusion and urgency ($p < 0.001$).

Model	Confusion Recall	Urgency Recall
SSMT-CU	**0.59**	**0.70**
ST-C	0.49	-
ST-U	-	0.59

Table 6: Performance evaluation for the subset of confused and urgent posts

These results indicate that by leveraging correlated tasks in the multi-task setting, the SSMT model has learned hidden abstractions which help it to outperform single-task models trained solely on confusion or urgency not just in general, but also in the more important subset of the data.

5. CONCLUSION

In this paper, we hypothesize that inter-label correlation or co-occurrence counts suggest but do not guarantee or quantify improvement in predictive performance with multi-task learning. This prompts us to design several experiments to explore the benefits of multi-task learning for confusion and urgency detection in MOOCs forums. We propose the SSMT model, a multi-task learning framework, to facilitate forum curation. We demonstrate that our proposed model outperforms single-task models consistently across both tasks. The multi-task framework takes advantage of the shared signals to yield not only superior performance in general, but also in the subset of the data that is most important for curation: posts that express both confusion and urgency. Future work can extend multi-task learning to explore its generalization performance across various course offerings. More specifically, it can investigate whether a multi-task learner trained on one course, can be effectively used for prediction in other related courses. In this regard, multi-task-based unsupervised domain adaptation can be applied to jointly learn the source and target course classifiers.

[4]We created this subset from test set of our data.

References

[1] Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. Youedu: Addressing confusion in MOOC discussion forums by recommending instructional video clips. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015*, pages 297–304.

[2] Muthu Kumar Chandrasekaran, Carrie Demmans Epp, Min-Yen Kan, and Diane J. Litman. 2017. Using discourse signals for robust instructor intervention prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3415–3421.

[3] Muthu Kumar Chandrasekaran, Min-Yen Kan, Bernard C. Y. Tan, and Kiruthika Ragupathi. 2015. Learning instructor intervention from MOOC forums: Early results and issues. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015*, pages 218–225.

[4] Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. 2014. Predicting instructor's intervention in MOOC forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1501–1511.

[5] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, pages 160–167.

[6] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJNLP)*, pages 1723–1732.

[7] Josh Gardner and Christopher Brooks. 2018. https://doi.org/10.1007/s11257-018-9203-z Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, 28(2):127–203.

[8] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

[9] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, page 2873–2879.

[10] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. https://doi.org/10.18653/v1/P17-1001 Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 1–10.

[11] Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley New York.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. https://doi.org/10.3115/v1/D14-1162 Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

[13] Antoine Pigeau, Olivier Aubert, and Yannick Prié. 2019. Success prediction in MOOCs: A case study. In *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019*, pages 390–395.

[14] Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé, and Lise Getoor. 2014. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, page 1272–1278.

[15] Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. 2014. https://doi.org/10.1145/2556325.2567879 Social factors that contribute to attrition in MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, page 197–198.

[16] Julia E Seaman, I Elaine Allen, and Jeff Seaman. 2018. Grade increase: Tracking distance education in the united states. *Babson Survey Research Group*.

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[18] Feng Wang and Li Chen. 2016. A nonlinear state space model for identifying at-risk students in open online courses. In *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, pages 527–532.

[19] Xu Wang, Diyi Yang, Miaomiao Wen, Kenneth R. Koedinger, and Carolyn Penstein Rosé. 2015. Investigating how student's cognitive behavior in MOOC discussion forum affect learning gains. In *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015*, pages 226–233.

[20] Xiaocong Wei, Hongfei Lin, Liang Yang, and Yuhai Yu. 2017. A convolution-lstm-based deep neural network for cross-domain MOOC forum post classification. *Information*, 8(3):92.

[21] Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 121–130.

[22] Ziheng Zeng, Snigdha Chaturvedi, and Suma Bhat. 2017. Learner affect through the looking glass: Characterization and detection of confusion in online courses. In *Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017*, pages 272–277.

[23] Ziheng Zeng, Snigdha Chaturvedi, Suma Bhat, and Dan Roth. 2019. https://doi.org/10.1145/3303772.3303810 DiAd: Domain adaptation for learning at scale. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge, LAK 2019*, pages 185–194.

[24] Yu Zhang, Ying Wei, and Qiang Yang. 2018. Learning to multitask. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5771–5782.

[25] Yu Zhang and Qiang Yang. 2017. http://arxiv.org/abs/1707.08114 A survey on multi-task learning. *CoRR*, abs/1707.08114.