

Article

A Convolution-LSTM-Based Deep Neural Network for Cross-Domain MOOC Forum Post Classification

Xiacong Wei ^{1,2,*} , Hongfei Lin ², Liang Yang ² and Yuhai Yu ² ¹ School of Software Engineering, Dalian University of Foreign Languages, Dalian 116044, China² School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China; hflin@dlut.edu.cn (H.L.); liang@dlut.edu.cn (L.Y.); yuyh@dlnu.edu.cn (Y.Y.)

* Correspondence: weixiacong@dlufl.edu.cn; Tel.: +86-411-8611-1252

Received: 1 June 2017; Accepted: 26 July 2017; Published: 30 July 2017

Abstract: Learners in a massive open online course often express feelings, exchange ideas and seek help by posting questions in discussion forums. Due to the very high learner-to-instructor ratios, it is unrealistic to expect instructors to adequately track the forums, find all of the issues that need resolution and understand their urgency and sentiment. In this paper, considering the biases among different courses, we propose a transfer learning framework based on a convolutional neural network and a long short-term memory model, called ConvL, to automatically identify whether a post expresses confusion, determine the urgency and classify the polarity of the sentiment. First, we learn the feature representation for each word by considering the local contextual feature via the convolution operation. Second, we learn the post representation from the features extracted through the convolution operation via the LSTM model, which considers the long-term temporal semantic relationships of features. Third, we investigate the possibility of transferring parameters from a model trained on one course to another course and the subsequent fine-tuning. Experiments on three real-world MOOC courses confirm the effectiveness of our framework. This work suggests that our model can potentially significantly increase the effectiveness of monitoring MOOC forums in real time.

Keywords: MOOC; cross-domain; transfer learning; classification; neural network

1. Introduction

Over the past six years, massive open online courses (MOOCs) have been increasingly used to deliver learning opportunities worldwide in a variety of domains. Between late 2011 and 2016, globally, over 58 million learners were enrolled in 6850 courses created by more than 700 institutions [1]. Learners often express feelings, exchange ideas and seek help by posting questions in discussion forums. These forums reflect learner affect, attitude and progress, which supply valuable feedback to teachers and education administrators. Instructors can better understand the gaps in learner knowledge or offer targeted feedback at a scale based on the discussion forum [2]. However, due to the very high learner-to-instructor ratios and the fact that hundreds of new discussion threads are created per day, it is unrealistic to expect instructors to adequately track the forums, find all of the issues that need resolution, understand their urgency and answer them individually. It is also difficult to recognize the sentiment in each post. As a result, instructors might overlook questions posed by struggling learners, thereby discouraging forum participation and decreasing learner motivation. A lack of responsiveness in forums might even result in a higher dropout rate. Finding posts that need resolution, understanding their urgency and recognizing the sentiment in such posts are all text classification tasks. If we could perform forum post classification in real time, the results would offer valuable insight to instructors for moderating and planning interventions within MOOC forums.

In many situations, we can collect certain forum posts from one course (i.e., the source domain), which is labeled, but we must predict various forum posts from different courses (i.e., the target domain), which are unlabeled. Different course forum posts have different feature spaces and distributions, and certain words may appear frequently in one course, but only sporadically or rarely in other courses. For example, the words “angle” and “delve” frequently appear in the *How_to_learn_Math* course, but never in the *SciWrite* course. If we directly apply a classification model trained on the source domain (course) to a target domain (course), the result is often unsatisfactory because the bias between the source and target domains hinders the learning of an accurate classification model. Publicly-labeled MOOC forum posts are scarce resources, and we cannot acquire labeled training data for every course. Moreover, manually annotating posts for a new course is an expensive and time-consuming task that requires linguists skilled in natural language processing (NLP). Thus, the question of how to use labeled forum posts from one course to predict forum posts in other courses is raised. Bakharia conducted some preliminary research on this issue [3] and determined that real-time classification of unlabeled data in a new course and the resulting low cross-domain classification accuracy highlight the need for transfer learning [4] algorithms.

Through observations, we find a few reasons for the low accuracy of cross-domain MOOC forum post classification attempts; these are the challenges that “off-the-shelf” NLP [5] transfer learning approaches should address when applied to educational data:

1. In a forum post, confusion/urgency/sentiment attitude are typically expressed in only one or two sentences, and most sentences do not express these factors. Therefore, a sentence in a positive post shows little difference from one in a negative post, leading to noisy data for classifiers.
2. The words/phrases used to express confusion/urgency are quite limited compared with other types of expressions. For example, posts frequently utilize phrases such as “What is”, “How will”, “Is there”, “Am I”, “How long” or “Where do we” to express confusion. Therefore, the set of usefully shared (or common) features in different domains is quite small.
3. In different domains, the methods used to express the same attitude are often quite similar. As a result, we encounter an imbalance problem; that is, the shared features are almost exclusively features indicating the positive class. Therefore, only the positive features are shared among different domains, whereas features indicating the negative class in different domains are highly diverse.
4. A post communicating confusion might be stated either explicitly or implicitly. Consider the following two real posts on an MOOC forum:
 Example 1. I understand that if you completed 65% you will get the certificate but where do we download the certificate?
 Example 2. I have tried to submit a document form of my response, but still nothing happens.
 Example 1 explicitly clearly expresses confusion related to the procedure to download the certificate. Example 2 is an example of implicit confusion. Recognizing explicit confusion is not difficult, whereas identifying implicit confusion requires deep semantic understanding.

In this paper, we address a novel problem that had not been previously addressed by the educational data mining (EDM) community: cross-domain MOOC forum post classification. We formulated confusion/sentiment/urgency identification as a binary classification problem and sought to develop a method that applies transfer learning to identify unlabeled MOOC forum posts from a new domain. Example posts that express confusion, urgency is it that be seen by an instructor and positive sentiment are displayed in the second row of Table 1, whereas the third row displays the opposite types of posts.

We conduct a preliminary transfer learning experiment involving the courses *How_to_learn_Math* (education (Edu)) and *SciWrite* (medicine (Med)). Following prior work on cross-domain classification, we rank the high-frequency features common to both domains in descending order and select the top 50 features as pivot features. For each pivot feature x_k , we select the feature w_s with the highest

co-occurrence in the source domain and the feature w_t with the highest co-occurrence in the target domain; these are referred to as non-pivot features. Thus, we construct a co-occurrence non-pivot feature pair (w_s, w_t) for pivot feature x_k . To reduce feature mismatches between the two courses, we transform every w_s that appears in the source domain with " $w_s w_t$ " and every w_t that appears in the target domain with " $w_s w_t$ ". In this manner, we build a common subspace between different courses. Finally, a support vector machine (SVM) classifier is trained on the transformed Edu to predict Med, and the accuracy increases by 1.2% compared with that achieved without transformation. Thus, if we can design a meaningful feature, it can improve the transfer learning accuracy. However, these conventional approaches require hand-crafted features that are empirical and task dependent. A given approach requires a considerable amount of engineering skill and domain expertise to construct features that are specific to a certain task. If we turn to another classification task, this feature engineering method might not work satisfactorily. However, this issue can be avoided if features can be learned automatically using a general-purpose learning procedure.

Table 1. Examples of confusion/urgency/sentiment posts.

| | Confusion | Urgency | Sentiment |
|-----|---|--|---|
| POS | I understand that if you completed 65% you will get the certificate but where do we download the certificate? | I hope any course staff member can help us to solve this confusion asap!!! | This is going to be an awesome course! |
| NEG | I have tried to submit a document form of my response, but still nothing happens. | I am Rana from Egypt, I study dentistry. Happy to be with you. Good luck for everyone. | I get frustrated & at times belittle myself thinking I am stupid :(|

Based on the special characteristics listed above, we determined that this problem is particularly well suited for deep neural networks. We propose a new transfer learning method based on the convolutional neural network (CNN) [6] and the long short-term memory (LSTM) [7] network, known as ConvL. We first apply the convolution operation to obtain a feature that is well depicted and then apply LSTM to learn a post representation, which considers the long-term temporal semantic relationships of features. We transfer the model parameters trained on the source domain course to the target domain course and subsequently fine-tune them. Experiments on a Stanford MOOC post dataset demonstrate that our proposed framework can effectively fulfill the tasks of classifying confusion/urgency/sentiment, learning a nondiscrimination feature representation from the source course and transferring it to the target course.

The major contributions of this work are as follows. First, we present the results obtained via the first study of cross-domain MOOC forum post classification. This method paves the way for instructor intervention within MOOC forums in real time. Second, EDM is a research field concerned with the application of data mining, machine learning and statistics to information generated in educational settings, and our study seeks to develop and improve methods for exploring these data and offers new insights related to EDM technology. Third, the proposed method promotes text transfer learning research.

The remainder of the paper is organized as follows. In Section 2, we review the literature on NLP in MOOC education research, deep learning in NLP and cross-domain transfer learning in text classification. Section 3 describes the problem and presents certain definitions. The methodology of this study is described in Section 4; the implementation and evaluation is described in Section 5; and the results and discussion are summarized in Section 6. Section 7 concludes our work and outlines future studies.

2. Related Works

2.1. NLP in MOOC Research

Much of the early related research on MOOC EDM has targeted the use of structured data. For example, Yang et al. applied a classification model using discussion forum behavior and clickstream data to automatically identify posts expressing confusion [8]. However, few studies have investigated the use of NLP in MOOC EDM. To predict course completion, Crossley et al. applied NLP tools, such as the Writing Assessment Tool (WAT), the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) and Tool for the Automatic Assessment of Sentiment (TAAS), to quantify forum post length, lexical sophistication, situation cohesion, cardinal numbers and trigram production. The results demonstrated that, when predicting MOOC completion, considering the writing quality is more useful than assessing observed behaviors [9]. Robinson et al. utilized unstructured language data to predict student success; their results demonstrated that an NLP-based model can predict completion better than one based on demographics. This approach emphasizes the potential for NLP in the prediction of student success [10]. Ramesh et al. proposed a model including both linguistic features (such as sentiment polarity and topic features) and behavioral features (such as lecture reviews, posting/voting/viewing discussion forum content) of MOOC discussion forums to predict student survival. The results demonstrated that linguistic features boost performance [11]. For the prediction of sentiment, Liu et al. proposed a feature selection method based on multi-swarm optimization to recognize the sentiment of online course reviews [12]. To understand learner performance, Tucker et al. quantified the sentiment of textual data expressed by students using NLP techniques to compute the sentiment of each word. The sum of the sentiment scores of every word in the text of a post was then used to reflect the sentiment of that post. The results can be used to understand the factors that influence student performance in MOOCs [13]. Wen et al. applied a simple sentiment analysis method to three MOOCs. Their results revealed a strong interrelationship between the sentiment of forum posts and the number of students who drop out [14]. To recognize confusion, Agrawal et al. used NLP to extract features from forum post text and metadata variables. Finally, a multiple classification model was implemented to detect confusion in forum posts. They also investigated confusion via NLP and used the results to retrieve the corresponding video clips [15].

These studies indicate that applying NLP techniques to unstructured language data can achieve better performance than studies on behavior data and also suggest a new way to improve MOOC research.

2.2. Deep Learning in NLP

In recent years, deep learning has achieved tremendous success in text mining. As a subfield of machine learning, deep learning aims to use learning-based methods to solve complex non-linear problems in a manner similar to the structure and processes of the human brain. By building a hierarchical structure, this approach can automatically capture useful intermediate feature representation using a general-purpose learning procedure, which is the key advantage of deep learning [16].

Further improvements in deep learning have been obtained with alternative types of neural network architectures, including the CNN and LSTM network. The CNN is a neural network that can learn the internal structure of data and performs well in text fields. The existing studies in which CNN has been applied to text have attempted to resolve many problems, such as sentence modeling [17], relational classification [18], sentence level text classification [19], machine translation [20] and domain adaptation mining of user consumption intention [21]. LSTM is a specific type of recurrent neural network (RNN) that has proven powerful for modeling long-range dependencies. The LSTM model and its many variants have achieved outstanding performance in sequence-learning problems involving text analysis [22,23]. All of these studies have verified the effectiveness of deep learning in NLP.

2.3. Cross-Domain Transfer Learning

Transfer learning is a method of resolving a lack of labeled data in cross-domain classification. It has been widely and successfully used in many domains, such as in the image field [24], human activity classification [25], software defect classification [26] and speech recognition [27].

In the text field, researchers have proposed various methods to tackle this problem based on non-deep neural networks [4,28–36]. Blitzer et al. presented structural correspondence learning (SCL) for transfer learning [4]. In this approach, an unlabeled dataset is used. Words that occur more frequently are considered as pivot features from the source, as well as target domain. Pan et al. proposed the spectral feature alignment (SFA) algorithm for cross-domain sentiment classification [28]. In SFA, a set of domain-independent sentiment words is identified at first. Then, the spectral clustering algorithm is adapted to co-cluster the domain-independent and the domain-specific features into a shared cluster space. Using these words the association between pivot and non-pivot features is estimated. Zhou et al. proposed a joint non-negative matrix factorization framework by linking heterogeneous input features with pivots for domain adaptation [29]. Bollegala et al. used a feature expansion method along with an automatically-constructed sentiment-sensitive thesaurus to train and test a binary classifier [30]. They also proposed a sentiment-sensitive word embedding learning method by constructing three objective functions: the distributional properties of pivots, label constraints in the source domain documents and geometric properties in the unlabeled documents in both the source and target domains. Xia et al. first proposed a labeling adaptation method using a parts-of-speech (POS)-based feature ensemble (FE) that assigns different weights to different POS tags, giving higher weights to domain-independent parts, such as adjectives and verbs, and lower weights to domain-specific parts, such as nouns [31]. They also presented a PCA-based sample selection (PCA-SS) method for instance adaptation. Combining FE with PCA-SS for domain adaptation results in significant improvements compared to either FE or PCA-SS alone. Huang et al. proposed a topic-related boosting-based learning framework named TR-TrAdaBoost for cross-domain sentiment classification. The idea of the model is to capture the latent semantic structure by extracting the topic distribution of documents, so as to embed both domain-specific and shared information of documents [32]. Li et al. propose a method named document concept vector (DCV) for the cross-domain text classification, which extracts the concept level information of the document [33]. Bhatt et al. presented a cross-domain classification method. It can learn an accurate model for the new unlabeled target domain given labeled data from multiple source domains where all domains have (possibly) different label sets [34]. Qu et al. proposed a transfer learning-based approach to named entity recognition in novel domains with label mismatch over a source domain [35]. Zoph et al. proposed a cross-language machine translation method. It transfers some of the learned parameters from a high resource language to initialize and constrain training the low-resource language [36].

Because deep neural networks are easily generalizable from one domain to another, the internal representation of the neural network contains no discrimination information on the raw input [37]. Recently, certain transfer learning methods based on deep neural networks have been proposed. Pan et al. reported a multi-layer transfer learning method based on non-negative matrix tri-factorization to address cross-domain text classification [38]. Collobert et al. proposed a framework capable of multi-task transfer learning [39]. Ding et al. presented a domain-adaptive framework based on the CNN for the identification of user consumption intention in the movie domain [21]. Wei et al. proposed a two-layer convolutional neural network (LM-CNN-LB) for cross-domain product review sentiment classification [40]. Various studies have also addressed learning representation via a deep architecture to reduce transfer loss [41–45]. To measure the transferability of the deep neural network in NLP, Mou et al. studied the transferability of semantic-relative and semantic-irrelative tasks, layers and parameter initialization in multi-task learning and their combination in three datasets [46].

However, to the best of our knowledge, no previous work has fully investigated the potential of deep learning in cross-domain MOOC forum post text classification.

3. Definitions

Domain: A domain D consists of two components: a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, \dots, x_n\} \in \chi$, n is the number of feature vectors in X , χ is the space of all possible feature vectors and X is a particular learning sample. In general, different domains always have different feature spaces or different marginal probability distributions. In our paper, a domain refers to a course. X refers to a forum post.

Source domain: $D_S = \{(X_{S_i}, Y_{S_i})\}_{i=1}^{n_S}$ refers to a set of labeled forum posts from a certain domain course, where X_{S_i} is the i -th labeled post. In our paper, this representation denotes one forum post in the source domain; Y_{S_i} is the label of X_{S_i} , $Y_{S_i} \in \{+1, -1\}$; and the labels $+1$ and -1 are positive and negative labels, respectively. In our paper, n_S denotes the number of labeled forum posts in the source domain.

Target domain: $D_T = \{(X_{T_i})\}_{i=1}^{n_T}$ refers to a set of unlabeled forum posts from another course, which is different from, but related to the source domain. X_{T_i} is the i -th unlabeled post, and in our paper, it denotes one forum post in the target domain. Additionally, n_T is the number of unlabeled forum posts in the target domain.

Cross-domain MOOC forum post classification: We define cross-domain MOOC forum post classification as the task of learning a binary classifier using D_S to predict the label Y_{T_i} of a post X_{T_i} in the target domain.

4. Methods

In this paper, we present a convolutional-LSTM neural network for cross-domain MOOC forum post text classification (ConvL). This method includes a word representation layer that transforms each word into a distributed input representation, a convolutional layer that extracts local contextual features and an LSTM layer that computes the representation of expressions by considering the semantic dependencies over a longer time. The parameters of the target domain model are first trained on the source domain and subsequently transferred to the target domain. We hold the embedding layer constant and allow the remaining layers to be fine-tuned with fewer target domain-labeled posts. The architecture of ConvL is illustrated in Figure 1.

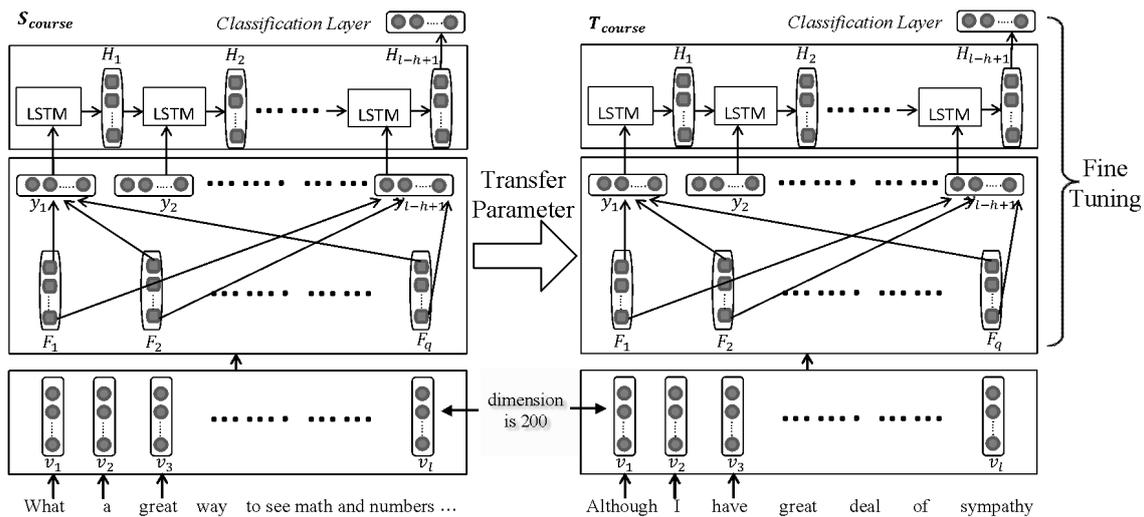


Figure 1. Architecture of convolutional-LSTM (ConvL) for cross-domain MOOC forum post text classification.

Because the CNN only works with fixed-length inputs, we set the length of every input to l by trimming longer sentences and padding shorter sentences with zeros. Given an input instance $X \in \mathbb{R}^l$, x_i is the i -th word in this instance.

Layer-0 input layer: The first layer is the embedding layer. By initializing the weight of the embedding layer with low-dimensional word vectors pre-trained by Word2Vec [47], each word x_i in the posts is mapped to its word vector $v_i \in \mathbb{R}^k$, where k is the dimension of the word vector. Thus, the post $X \in \mathbb{R}^{l \times k}$ is concatenated with the word vector as follows:

$$X_{1:l} = v_1 \oplus v_2 \oplus \dots \oplus v_l \tag{1}$$

The word vector is fine-tuned during training in the source domain. Subsequently, we regularize our network via dropout, and the output can be fed to a further neural network layer.

Layer-1 convolutional layer: This layer consists of a one-dimensional convolution operation. The essential role of the convolutional layer is to convert text regions of a fixed size into feature vectors. The vector $m \in \mathbb{R}^{h \times k}$ is the filter of the convolution (i.e., the weight). The number of filters is p . The filter width h enables the convolutional layer to make use of the text word order to capture the contextual feature of a word. For a word vector v_i , we first concatenated the feature vectors around v_i within h and fed it to the convolution operation to take the dot product of the weight vector $m \in \mathbb{R}^{h \times k}$ and the vector of inputs $v_i \in \mathbb{R}^k$ with the activation function. A new feature representation of the i -th word in a post was thus created:

$$f_i = \sigma(\mathbf{m}^T v_{i:i+h-1} + b) \tag{2}$$

where σ is a non-linear activation function. The weight vector $\mathbf{m} \in \mathbb{R}^{h \times k}$ and bias b are shared by all units in the same layer and are learned through training. Using the filter $\mathbf{m} \in \mathbb{R}^{h \times k}$, the feature representation of the post $F \in \mathbb{R}^{l-h+1}$ was created as follows:

$$F = [f_1, f_2, \dots, f_{l-h+1}] \tag{3}$$

The output of this layer $y_i \in \mathbb{R}^{(l-h+1) \times q}$ was concatenated with the i -th dimension of every feature map as follows:

$$y_i = F_1^i \oplus F_2^i \oplus \dots \oplus F_{l-h+1}^i \tag{4}$$

Layer-2 LSTM layer: We took the output of the convolutional layer as the input to this layer. The LSTM model contains multiple LSTM cells, where each LSTM cell consists of an input gate (i_t), forget gate (f_t) and output gate (o_t). The calculation is as follows:

$$i_t = \sigma(W_i \cdot [H_{t-1}, y_t] + b_i) \tag{5}$$

$$f_t = \sigma(W_f \cdot [H_{t-1}, y_t] + b_f) \tag{6}$$

$$o_t = \sigma(W_o \cdot [H_{t-1}, y_t] + b_o) \tag{7}$$

$$g_t = \tanh(W_g \cdot [H_{t-1}, y_t] + b_g) \tag{8}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t \tag{9}$$

$$H_t = o_t \odot \tanh(C_t) \tag{10}$$

where \odot represents element-wise multiplication, σ denotes the sigmoid function containing the gating values in $[0,1]$. $y_t \in \mathbb{R}^q$ is the current input from the lower layer at time step t (where q is the dimensionality of word vector y_t). H_{t-1} is the output vector of the previous step, and $W_i, b_i, W_f, b_f, W_o, b_o, W_g$ and b_g are the parameters that must be trained. The last hidden vector is treated

as the post representation. Subsequently, we regularized our network via dropout, and the output can be fed to a further neural network layer.

Layer-3 classification layer: The output of the LSTM layer is fed into a fully-connected layer for classification.

In this scheme, the CNN-LSTM network was obtained and trained on the source domain data. The model obtained is denoted as S_{course} . To perform the transfer, we built another model (denoted as T_{course}) for which the neural network architecture is the same as that of S_{course} . We transferred all of the parameters from S_{course} to initialize the corresponding layers of T_{course} . We froze the embedding layer of T_{course} , and the parameters of the remaining layers were fine-tuned on a small quantity of target domain-labeled data. The model T_{course} thus obtained was used to predict the target domain. The first claim we make is that S_{course} can automatically learn feature representation that can be shared across the source and target domains and is beneficial for our classification tasks. The second claim we make is that fine-tuning the parameters on fewer target domain data transferred from S_{course} can enhance T_{course} such that it can learn features specific to the target domain classification task and further boost accuracy. The third claim we make is that CNN can effectively use local contextual features to capture the generic factors of variation; LSTM can learn a feature representation by considering the semantic long-term dependency. The combination is better than the CNN-based and LSTM-based network separately.

5. Implementation and Evaluation

5.1. Dataset

The Stanford MOOC posts dataset [48] contains forum posts pertaining to three domain areas: humanities/sciences (HS), medicine (Med) and education (Edu). HS contains 6 courses: GlobalHealth/WomensHealth/Winter2014 (2254 posts), HumanitiesScience/StatLearning/Winter2014 (3112 posts), HumanitiesScience/Stats216/Winter2014 (341 posts), HumanitiesSciences/EP101/Environmental_Physiology (2549 posts), HumanitiesSciences/Econ-1/Summer2014 (1584 posts) and HumanitiesSciences/Econ1V/Summer2014 (160 posts). Med contains 4 courses: Medicine/HRP258/Statistics_in_Medicine (3321 posts), Medicine/MedStats/Summer2014 (1218 posts), Medicine/SciWrite/Fall2013 (5184 posts), Medicine/SURG210/Managing_Emergencies_What_Every_Doctor_Must_Know (279 posts). Edu contains one course: Education/EDUC115N/How_to_Learn_Math (10,000 posts). Each post was manually annotated. Similar to [3], we selected the courses of each domain with the largest amount of forum posts for evaluation. Only three tasks—confusion, urgency and sentiment—and their classifications were used in our experiment. Table 2 displays the names of the courses, the number of posts and the size of each category evaluated in our experiment. We framed the problem of detecting confusion/urgency/sentiment as a binary classification task. Posts with a confusion/urgency/sentiment rating greater than four in the MOOC post dataset fall into the “confused/urgency/sentiment” class (pos); all other posts fall into the “not confused/urgency/sentiment” class (neg). In each classification task, we constructed 6 cross-domain classification subtasks—Edu → Med, Edu → HS, Med → Edu, Med → HS, HS → Edu and HS → Med—in which the letters preceding the arrow correspond to the source domain, and the letters after the arrow correspond to the target domain.

Table 2. Evaluation dataset for cross-domain MOOC forum post classification.

| Domain | Size | Confusion | | Urgency | | Sentiment | |
|----------------------------------|------|-----------|-----|---------|-----|-----------|-----|
| | | POS | NEG | POS | NEG | POS | NEG |
| EDUC115N/How_to_Learn_Math (Edu) | 9879 | 32% | 68% | 5% | 95% | 83% | 17% |
| SciWrite/Fall2013 (Med) | 5184 | 91% | 9% | 38% | 62% | 75% | 25% |
| StatLearning/Winter2014 (HS) | 3030 | 84% | 16% | 32% | 68% | 94% | 6% |

5.2. Experimental Setup

During preprocessing, punctuation, hyperlinks and signs replaced by automated anonymization (e.g., “*<phoneRedac>*”, “*<ZipRedac>*”) were removed, and all characters were converted to lower case. The weight of the input layer in ConvL was initialized with 200-dimensional word vectors. The word vectors were pre-trained with the Stanford MOOC post dataset using Google Word2Vec (<https://code.google.com/p/word2vec/>). The training parameters are as follows: using Skip-Gram model (cbow is 0), size is 200, window is 8, negative is 25, using HS method (hs is 0), sample is 1e-4, threads is 20, binary is 1 and iter is 15. The corpus contained 29,604 forum posts and 1.03 million words, and the vocabulary totaled 20,122 words. The neural network architecture we proposed was implemented in Keras 2.0 [49]. In the convolutional layer, for computational reasons, the length of every input for review was 500; the number of convolution filters was 250; the filter width was 5; the border mode was valid; and the activation function was ReLU. In the LSTM layer, the dimensions of the hidden states and cell states in the LSTM cells were both set to 100. In the classification layer, the activation function of the fully connected layer (L2 regularizers) was sigmoid, and the number of hidden units was 1. The batch size of the neural network was 32, and the optimizer was rmsprop. Dropout training is a method of regularizing the model by randomly leaving out features, and in our experiment, all dropout values were set to 0.25. Shuffling was performed after every epoch. The training procedure periodically evaluated the binary cross-entropy objective function on the training and validation sets. The S_{course} network was trained for 3 epochs, and the T_{course} network was trained for 50 epochs. The evaluation metric was accuracy. The test performance was associated with the validation accuracy of the last epoch. In the process of training S_{course} for ConvL, the training set comprised 9/10 of the source domain data, and the validation set consisted of the remaining 1/10. In the process of training T_{course} , the training and validation sets were each 1/10 of the target domain, and the remainder constituted the test set. Our networks are trained on one NVIDIA Tesla K20c GPU in a 64-bit Dell computer with two 2.40-GHz CPUs, 64 G main memories in Dalian, China, and Ubuntu 12.04. For example, in the cross-domain classification subtask, Edu → Med, one epoch requires 9233 s when training S_{course} on Edu and 136 s when training T_{course} on Med.

5.3. Baselines

To evaluate the effectiveness of ConvL, we compared our proposed method with several existing algorithms:

- CNN-NTL: This method is similar to the CNN-TL method mentioned above, but we used S_{course} to directly predict the unlabeled posts in the target domain without using a transfer scheme.
- CNN-TL: This method uses a model and transfer scheme similar to ConvL, but the layer subsequent to the convolutional layer is a max pooling layer instead of an LSTM layer. The max pooling length is 2.
- LSTM-NTL: This method is similar to the LSTM-TL mentioned above, but we used S_{course} to directly predict the unlabeled posts in the target domain without using a transfer scheme.
- LSTM-TL: This is a model transfer scheme similar to ConvL, but it has no convolutional layer.
- Consumption intention mining model (CIMM)-TL: CIMM is a CNN-based framework proposed in [21] for mining consumption intention. It consists of a convolutional layer, a max pooling layer and two fully-connected layers. In their study, the authors investigated the possibility of transferring the CIMM mid-level sentence representation learned from one domain to another by adding an adaptation layer. We refer to this method as CIMM-TL. Because framework always performs worse without using transfer learning schema, we do not display CIMM-NTL.
- LM-CNN-LB: LM-CNN-LB is a two-layer convolutional neural network for cross-domain product review sentiment classification proposed in [40].
- ConvL-NTL: This method is the same as ConvL, but we used S_{course} to directly predict the unlabeled posts in the target domain without using a transfer scheme.

- ConvL: This is the method proposed in this paper. The code is publicly available (<https://github.com/wxcmm/NLP>).
- ConvL-in domain: This shows the ConvL results of the validation set for training on the source domain. The result can be treated as a higher bound.

6. Results and Discussion

The performances of the different methods on each task are shown in Table 3. CNN-NTL can effectively use local contextual features (via the convolution operation) and global contextual features (via the max pooling operation) to capture the generic factors of variation; thus, it can be treated as a feature extractor. LSTM-NTL can learn a feature representation by considering the semantic long-term dependency. All of these methods can automatically capture a useful intermediate feature representation for cross-domain text classification. When we applied the transfer scheme proposed in this paper, CNN-TL and LSTM-TL performed better than CNN-NTL and LSTM-NTL. By training S_{course} on the source domain course forum posts, the methods could automatically extract nondiscriminatory feature representations that can be shared across the source and target domains. Fine-tuning the parameters transferred from S_{course} using only small amounts of target domain data adapts T_{course} to learn features that are specific to the target domain task.

Table 3. Experimental results from different methods (accuracy %). CIMM, consumption intention mining model.

| Method | Confusion | Sentiment | Urgency |
|-----------------|--------------|--------------|--------------|
| CNN-NTL | 63.35 | 83.52 | 65.90 |
| LSTM-NTL | 62.23 | 83.30 | 78.95 |
| CNN-TL | 81.89 | 83.74 | 79.72 |
| LSTM-TL | 82.25 | 85.88 | 86.13 |
| CIMM-TL | 68.87 | 84.08 | 24.24 |
| LM-CNN-LB | 78.46 | 81.83 | 84.59 |
| ConvL-NTL | 67.52 | 80.91 | 79.99 |
| ConvL | 81.45t | 85.91 | 86.69 |
| ConvL-in domain | 81.88 | 86.08 | 89.14 |

CNNs and LSTMs are individually limited in their modeling capabilities. The CNN-TL can obtain features that are well depicted through the convolution operation, but the max pooling operation leads to a loss of contextual information. It may fail to capture long-distance dependency. LSTMs is the temporal modeling that is done on the input feature. It can address the limitation of CNN by sequentially modeling texts across sentences. However, higher level modeling of input feature can help to disentangle underlying factors of variation within the input, which should then make it easier to learn temporal structure between successive time steps. Thus, in our work, we take advantage of the complementarity of CNNs and LSTMs by combining them into one unified architecture. LSTM performance can be improved by providing better features to the LSTM, which the CNN layers provide through the convolution operation. Using the transfer parameters from S_{course} trained on the source domain to initialize T_{course} and then fine-tuning with smaller amounts of target domain data, our method can learn feature representations that are generalizable across different domains. Consequently, ConvL outperforms all of the baseline methods in sentiment and urgency classification tasks. We also observe that when different classification tasks are available, the best learning strategy varies. For confusion classification, ConvL is less effective than the LSTM-based method; CIMM-TL and LM-CNN-LB are not competitive in any of the three tasks investigated here. This is probably because CIMM-TL and LM-CNN-LB are CNN-based networks. They do not consider long-term temporal semantic relations of words. Additionally, for CIMM-TL, not initializing the last fully-connected layer with transferring parameters and freezing too many layers lead to poor performance.

Below, we analyze various factors that can influence the effectiveness of our method.

Corpus for training word vector: We compared the word-embedding weight pre-trained by Google News and the MOOC corpus. The dataset is as depicted in Section 5.1 and the experimental parameters as depicted in Section 5.2, except the word embedding and the baseline model is ConvL. Figure 2 shows that as a weight for the embedding layer, the word vector pre-trained by the MOOC corpus results in performance superior to that of the word vector pre-trained by Google News on all tasks. We conclude that, for cross-domain MOOC forum post classification, a word vector pre-trained using a corpus specific to the task performs better and can introduce more relevant semantic knowledge than one pre-trained on Google News.

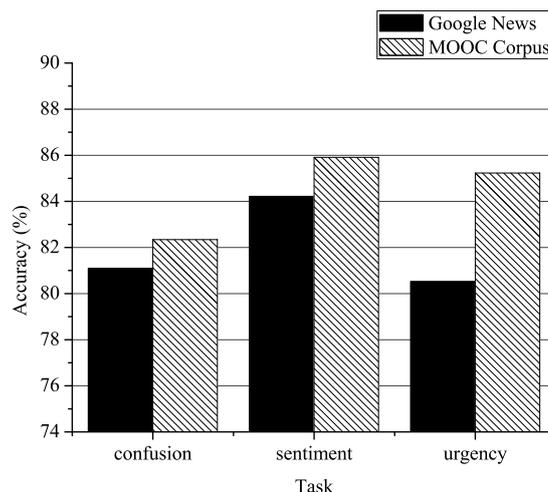


Figure 2. Comparison of word embeddings pre-trained on Google News or on the MOOC corpus.

Epoch number for training on the source domain: We also analyzed the effect of the epoch number during the training of S_{course} . The dataset is as depicted in Section 5.1 and experimental parameters as depicted in Section 5.2, except the epoch number during the training of S_{course} and the baseline model is ConvL. As shown in Figure 3, the accuracy increases for Epochs 1–3, but as the number of epochs increases further, the performance does not continue to improve. This result suggests that parameters that are well trained using the source domain are crucial for our work and that S_{course} learns a discriminative feature representation that can satisfactorily initialize T_{course} . However, if we train excessively on the source domain, the parameters transferred from S_{course} to T_{course} become overfitted to the source domain, leading to a situation in which the learned feature representation of S_{course} is excessively specific to the source domain and, thus, underfits the target domain.

Time to readiness for transfer: In this subsection, we evaluate the transferability of ConvL by freezing different layers of T_{course} as shown in Figure 4. The dataset is as depicted in Section 5.1 and experimental parameters as depicted in Section 5.2, except freezing layers and the baseline model is ConvL. For example, we froze the input-layer of T_{course} after transferring the parameters from S_{course} . The parameters of the layers (i.e., the convolutional layer, LSTM layer and classification layer) were trainable and fine-tuned on only small numbers of labeled reviews of the target domain. Figure 4 shows that the fine-tuning of additional layers stimulates learning features that are specific to the target domain. When more layers participate in the fine-tuning, the features will be more specific to the target domain, and higher accuracy will be achieved.

Labeled target domain data: We also experimented with different proportions of labeled target domain data by varying them from 1/20–1/2. The dataset is as depicted in Section 5.1 and experimental parameters as depicted in Section 5.2, except the proportions of labeled target domain for fine-tuning and the baseline model is ConvL. According to Table 4, each row refers to the average accuracy of

all six cross-domain classification subtasks. The results clearly indicate that the use of the target domain labeled information is critical for learning effective features for transfer learning. This result indicates that our framework learns information that is more specific to the target domain and is more meaningful for cross-domain MOOC forum post classification.

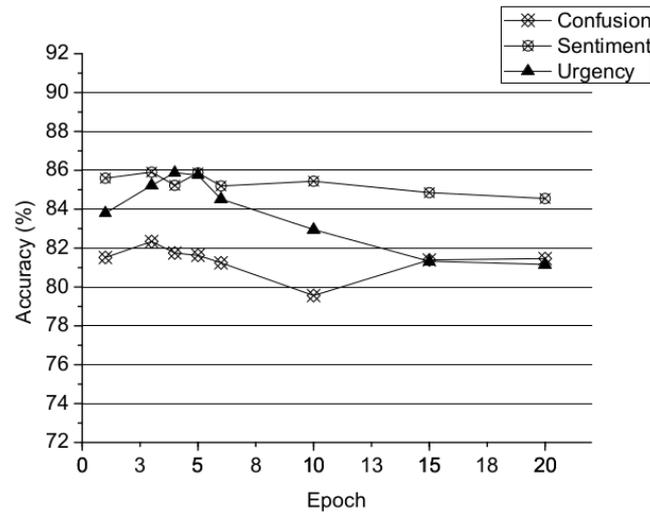


Figure 3. Effect of the S_{course} epoch number.

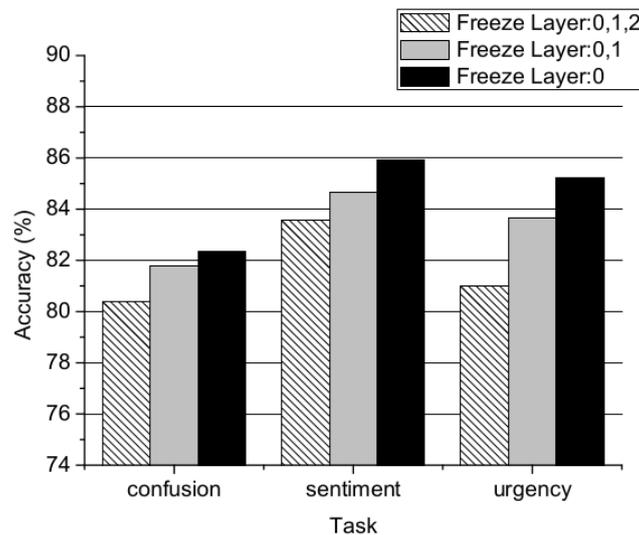


Figure 4. Comparison of the accuracy of ConvL when different layers are frozen.

Table 4. Effect of the labeled target domain data ratio for training T_{course} (average accuracy of all cross-domain classification subtasks %).

| Ratio of Target Domain Training Data | Accuracy (%) |
|--------------------------------------|--------------|
| 1/20 | 82.09 |
| 2/20 | 84.49 |
| 3/20 | 84.54 |
| 4/20 | 84.68 |
| ... | ... |
| 10/20 | 85.93 |

Multi-source course: We also compared multi-source and single-source domains. The data are from the dataset as depicted in Section 5.1, and the baseline model is ConvL. For the multi-source domain, we used one domain as the target domain and the remaining two domains as the source domain. We constructed three cross-domain classification subtasks: {Edu, HS} → Med, {Edu, Med} → HS, {Med, HS} → Edu. In the process of training S_{course} , the training set consisted of 9/10 of the two source domain data, and the validation set was the remaining 1/10. The process of training T_{course} was the same. The rest of the parameters are as depicted in Section 5.2. Deep learning has been demonstrated to work well on large corpuses. As shown in Figure 5, for urgency classification, as the source domain training data expands, more useful knowledge is introduced, which is advantageous for cross-domain urgency classification. The accuracy of training on multi-source domains was 86.55%. This performance exceeds even that of LSTM-TL trained on one source domain. However, for the confusion/sentiment classification task, the performance of our method when using a multi-source domain is worse. Thus, different classification tasks require different training sample selection methods. Cross-domain performance varies significantly depending on the choice of training samples.

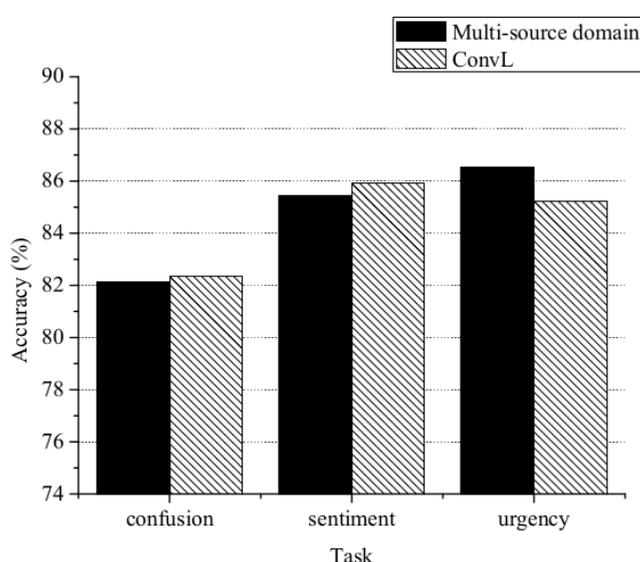


Figure 5. Comparison of a multi-source domain course and ConvL.

7. Conclusions

In this paper, we report the first attempt to apply transfer learning to confusion, urgency and sentiment detection in MOOC forum posts. We propose a deep learning framework based on convolution and LSTM that can capture the generic factors of variation present in all of the factors suitable for our three cross-domain MOOC forum post classification tasks. Our framework avoids feature engineering and can perform automatically. The results suggest that vectors pre-trained on an MOOC corpus are better than those pre-trained on Google News for cross-domain MOOC forum post classification. Using an appropriate number of epochs for training on the source domain results in better performance, while freezing excessive layers negatively impacts the transfer performance. Training with additional labeled data from the target domain for fine-tuning leads to learning information that is more specific to the target domain. The multi-source approach is a valid strategy, but depends on the choice of training sample selection.

In the future, we plan to study a combination of data selection and transfer learning approaches that complement each other and use labeled multi-source domain data to maximize their effectiveness. Research on text transfer learning based on deep neural networks and NLP in MOOCs is only in its infancy, and additional efforts are needed to resolve the various issues faced by MOOCs using NLP.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (No. 61632011, No. 61572102 and No. 61562080), the Natural Science Foundation of Liaoning Province (No. 20170540231), the National Social Science Foundation of China (No. 15BYY028) and Dalian University of Foreign Languages Research Foundation (No. 2014XJQN14 and No. 2014XJQN15).

Author Contributions: X.W. designed and wrote the paper; H.L. supervised the work; X.W. and Y.Y performed the experiments; and Y.L. analyzed the data. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|------|------------------------------|
| MOOC | Massive open online course |
| CNN | Convolutional neural network |
| LSTM | Long short-term memory |
| NLP | Natural language processing |

References

1. Shah, D. Monetization Over Massiveness: Breaking Down MOOCs by the Numbers in 2016. EdSurge. Available online: <https://www.edsurge.com/> (accessed on 25 July 2017).
2. Rossi, L.A.; Gnawali, O. Language independent analysis and classification of discussion threads in Coursera MOOC forums. In Proceedings of the Information Reuse and Integration, Redwood City, CA, USA, 13–15 August 2014; pp. 654–661.
3. Bakharia, A. Towards Cross-domain MOOC Forum Post Classification. In Proceedings of the L@S: ACM Conference on Learning at Scale, Edinburgh, Scotland, UK, 25–26 April 2016; pp. 253–256.
4. Blitzer, J.; McDonald, R.; Pereira, F. Domain adaptation with structural correspondence learning. In Proceedings of the Empirical Methods on Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 120–128.
5. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2000; pp. 1–1024.
6. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
7. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
8. Yang, D.; Wen, M.; Howley, I.; Kraut, R.; Rosé, C. Exploring the effect of confusion in discussion forums of massive open online courses. In Proceedings of the L@S: ACM Conference on Learning at Scale, Vancouver, BC, Canada, 14–18 March 2015; pp. 121–130.
9. Crossley, S.; McNamara, D.S.; Baker, R.; Wang, Y.; Paquette, L.; Barnes, T.; Bergner, Y. Language to Completion: Success in an Educational Data Mining Massive Open Online Class. In Proceedings of the International Conference on Educational Data Mining, Madrid, Spain, 26–29 June 2015.
10. Robinson, C.; Yeomans, M.; Reich, J.; Hulleman, C.; Gehlbach, H. Forecasting student achievement in MOOCs with natural language processing. In Proceedings of the Conference on Learning Analytics & Knowledge, Edinburgh, UK, 25–29 April 2016; pp. 383–387.
11. Ramesh, A.; Goldwasser, D.; Huang, B.; Daumé, D., III; Getoor, L. Understanding MOOC discussion forums using seeded LDA. In Proceedings of the Innovative Use of NLP for Building Educational Applications Conference, Baltimore, MD, USA, 22–27 June 2014; pp. 28–33.
12. Liu, Z.; Liu, S.; Liu, L.; Sun, J.; Peng, X.; Wang, T. Sentiment recognition of online course reviews using multi-swarm optimization-based selected features. *Neurocomputing* **2016**, *185*, 11–20.
13. Tucker, C.S.; Dickens, B.; Divinsky, A. Knowledge Discovery of Student Sentiments in MOOCs and Their Impact on Course Performance. In Proceedings of the International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Buffalo, New York, NY, USA, 17–20 August 2014; p. V003T04A0288.

14. Wen, M.; Yang, D.; Rosé, C.P. Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In Proceedings of the International Conference on Educational Data Mining, London, UK, 4–7 July 2014.
15. Agrawal, A.; Venkatraman, J.; Leonard, S.; Paepcke, A. YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In Proceedings of the International Conference on Educational Data Mining, Madrid, Spain, 26–29 June 2015.
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2016**, *521*, 436–444.
17. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. In Proceedings of the Association for Computational Linguistics Conference, Baltimore, MD, USA, 22–27 June 2014; pp. 655–665.
18. Nguyen, T.H.; Grishman, R. Relation extraction: Perspective from con-volutional neural networks. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 39–48.
19. Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the Conference on Empirical Methods on Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
20. Meng, F.; Lu, Z.; Wang, M.; Li, H.; Jiang, W.; Liu, Q. Encoding source language with convolutional neural network for machine translation. In Proceedings of the Association for Computational Linguistics Conference, Beijing, China, 26–31 July 2015; pp. 20–30.
21. Ding, X.; Liu, T.; Duan, J.; Nie, J.Y. Mining user consumption intention from social media using domain adaptive convolutional neural network. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Austin, TX, USA, 25–30 January 2015; pp. 2389–2395.
22. Chen, H.; Sun, M.; Tu, C.; Lin, Y.; Liu, Z. Neural sentiment classification with user and product attention. In Proceedings of the Conference on Empirical Methods on Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 1650–1659.
23. Tang, D.; Qin, B.; Feng, X.; Liu, T. Effective LSTMs for Target-Dependent Sentiment Classification. In Proceedings of the International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016; pp. 3298–3307.
24. Kandaswamy, C.; Silva, L.M.; Alexandre, L.A.; Santos, J.M.; de Sá, J.M. Improving deep neural network performance by reusing features trained with transductive transference. In Proceedings of the International Conference on Artificial Neural Networks. A conference of the European Neural Network Society, Hamburg, Germany, 15–19 September 2014; pp. 265–272.
25. Harel, M.; Mannor, S. Learning from multiple outlooks. In Proceedings of the International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 401–408.
26. Nam, J.; Kim, S. Heterogeneous defect prediction. In Proceedings of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Bergamo, Italy, 30 August–4 September 2015; pp. 508–519.
27. Huang, J.T.; Li, J.Y.; Yu, D.; Deng, L.; Gong, Y.F. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7304–7308.
28. Pan, S.J.; Ni, X.; Sun, J.T.; Yang, Q.; Chen, Z. Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the International World Wide Web Conference, Raleigh, WA, USA, 26–30 April 2010; pp. 751–760.
29. Zhou, G.; He, T.; Wu, W.; Hu, X.T. Linking heterogeneous input features with pivots for domain adaptation. In Proceedings of the International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 1419–1425.
30. Bollegala, D.; Mu, T.; Goulermas, J.Y. Cross-Domain Sentiment Classification Using Sentiment Sensitive Embeddings. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 398–410.
31. Xia, R.; Zong, C.; Hu, X.; Cambria, E. Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *Intell. Syst.* **2013**, *28*, 10–18.
32. Huang, X.; Rao, Y.; Xie, H.; Wong, T.; Wang, F. Cross-Domain Sentiment Classification via Topic-Related TrAdaBoost. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference, San Francisco, CA, USA, 4–9 February 2017; pp. 4939–4940.

33. Li, Y.; Wei, B.; Yao, L.; Chen, H.; Li, Z. Knowledge-based document embedding for cross-domain text classification. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 1395–1402.
34. Bhatt, H.S.; Sinha, M.; Roy, S. Cross-domain Text Classification with Multiple Domains and Disparate Label Sets. In Proceedings of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1641–1650.
35. Qu, L.; Ferraro, G.; Zhou, L.; Hou, W.; Baldwin, T. Named Entity Recognition for Novel Types by Transfer Learning. In Proceedings of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 899–905.
36. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer learning for low-resource neural machine translation. In Proceedings of the Empirical Methods on Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 1568–1575.
37. Lu, J.; Behbood, V.; Hao, P.; Xue, S.; Zhang, G. Transfer learning using computational intelligence: A survey. *Knowl. Based Syst.* **2015**, *80*, 14–23.
38. Pan, J.; Hu, X.; Li, P.; Li, H.; Li, W.; He, Y.; Zhang, Y.; Lin, Y. Domain adaptation via multi-layer transfer learning. *Neurocomputing* **2016**, *190*, 10–24.
39. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
40. Wei, X.C.; Lin, H.F.; Yu, Y.H.; Yang, L. Low-Resource cross-Domain product review sentiment classification based on a CNN with an auxiliary large-Scale corpus. *Algorithms*. **2017**, *10*, 81.
41. Glorot, X.; Bordes, A.; Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In Proceedings of the International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 513–520.
42. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In Proceedings of the International Conference on Machine Learning, Edinburgh, Scotland, UK, 26 June–1 July 2012; pp. 17–36.
43. Mesnil, G.; Dauphin, Y.; Glorot, X.; Rifai, S.; Bengio, Y.; Goodfellow, I.J.; Lavoie, E.; Muller, X.; Desjardins, G.; Warde-Farley, D. Unsupervised and Transfer Learning Challenge: a Deep Learning Approach. In Proceedings of the International Conference on Machine Learning, Scotland, UK, 26 June–1 July 2012; pp. 97–110.
44. Liu, B.; Huang, M.; Sun, J.; Zhu, X. Incorporating domain and sentiment supervision in representation learning for domain adaptation. In Proceedings of the International Congress On Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 1277–1283.
45. Gani, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2015**, *17*, 1–35.
46. Mou, L.; Meng, Z.; Yan, R.; Li, G.; Xu, Y.; Zhang, L.; Jin, Z. How Transferable are Neural Networks in NLP Applications? In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; pp. 479–489.
47. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
48. The Stanford MOOCPosts Data Set. Available online: <http://datastage.stanford.edu/StanfordMoocPosts/> (accessed on 17 March 2012).
49. Keras, C.F. GitHub. Available online: <http://github.com/fchollet/keras> (accessed on 25 July 2017).

