Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs

ZENUN KASTRATI[ID][1], **ALI SHARIQ IMRAN**[ID][2], **(Member, IEEE), AND ARIANIT KURTI**[1]
[1]Department of Computer Science and Media Technology, Linnaeus University, 351 95 Växjö, Sweden
[2]Department of Computer Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Corresponding author: Zenun Kastrati (zenun.kastrati@lnu.se)

ABSTRACT Students' feedback is an effective mechanism that provides valuable insights about teaching-learning process. Handling opinions of students expressed in reviews is a quite labour-intensive and tedious task as it is typically performed manually by the human intervention. While this task may be viable for small-scale courses that involve just a few students' feedback, it is unpractical for large-scale cases as it applies to online courses in general, and MOOCs, in particular. Therefore, to address this issue, we propose in this paper a framework to automatically analyzing opinions of students expressed in reviews. Specifically, the framework relies on aspect-level sentiment analysis and aims to automatically identify sentiment or opinion polarity expressed towards a given aspect related to the MOOC. The proposed framework takes advantage of weakly supervised annotation of MOOC-related aspects and propagates the weak supervision signal to effectively identify the aspect categories discussed in the unlabeled students' reviews. Consequently, it significantly reduces the need for manually annotated data which is the main bottleneck for all deep learning techniques. A large-scale real-world education dataset containing around 105k students' reviews collected from Coursera and a dataset comprising of 5989 students' feedback in traditional classroom settings are used to perform experiments. The experimental results indicate that our proposed framework attains inspiring performance with respect to both the aspect category identification and the aspect sentiment classification. Moreover, the results suggest that the framework leads to more accurate results than the expensive and labour-intensive sentiment analysis techniques relying heavily on manually labelled data.

INDEX TERMS Aspect identification, aspect sentiment classification, course evaluation, CNN, data mining, deep learning, eLearning, LSTM, MOOC, opinion mining, online course reviews, sentiment analysis, students' feedback, weak supervision.

I. INTRODUCTION

The advances in the technologies has brought a lot of innovation into the education field. One of those undoubtedly are Massive Open Online Courses-MOOCs that researchers refer to as the fourth stage of evolution of the online education [1]. MOOCs are considered also to be the ultimate way for educational content delivery to the most distanced students. MOOCs became even more actual especially nowadays, when the entire world is fighting against global COVID-19 pandemic, which forced lots of universities to either close or move completely their operations on the distance mode [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Pengcheng Liu[ID].

Despite the current increase of interest in MOOCs globally, they still suffer from high dropout rates [3]. The research indicates that the main factors that influence the retention rates in MOOC-like courses are closely connected to course design and learners' behaviour [4].

Motivated by these trends and as a part of a project that focused on a co-creation process with IT industry, we developed a number of courses trying to investigate the increased flexibility, teaching approaches, and the use of educational technologies for busy IT professionals [5]. Beside co-creation in the analyses and the design phase of the courses, our aim was to also make active use of students' feedback in order to reach a continuous improvement of the course content and delivery for eLearning [6] and MOOC classification

frameworks [7]. Students' feedback helps us to implement a co-creation process throughout the project life-cycle. In this aspect we have conducted a minor pilot study with one of the MOOC-like courses developed in cooperation with the industry. The initial results from this pilot indicated that the course structure and content followed by lecturer-student interactions constitute the three substantive aspects that need to be taken into account for an effective online class [8]. Based on this, our aim with the research efforts reported in this study is to investigate the main aspects that are most important on MOOC effectiveness by making use of students' reviews available in Coursera.

Manually analyzing each student's review would be extremely challenging due to large-scale feedback of MOOCs. On the contrary, automatic techniques including conventional machine learning algorithms and deep neural networks have proven to be useful in this context. However, they require very large amount of labelled data for training the model, which in many domains, especially education, are either not readily available or are costly to obtain because of intensive human efforts. Therefore, in this study we present a framework to automatically examine qualitative opinions of students expressed in reviews by eliminating the need for most, if not all, human interventions. The proposed framework takes advantage of the weak supervision strategy, which aims to train a deep network to automatically identify the aspects that are most important on MOOC effectiveness using either very few or even no manual annotations. To achieve this, a weak supervision signal constituted by either a small amount of labeled reviews per aspect category or a set of keywords associated to them is used for learning model parameters. The trained parameters are then fed into the deep neural network to predict the aspect categories for unlabeled reviews. In addition, the framework examines the attitude that the students expressed towards the aspects commented on a given review.

The proposed framework is tested on two real-life datasets containing students' feedback of courses offered using two delivery modes, online-delivery mode and traditional face-to-face delivery mode.

MOOCs are multidimensional constructs where the teaching effectiveness is influenced by various factors including those related to creation, execution, and evaluation [9]. The ADDIE instructional design process, which has found wide acceptance and use from MOOC instructors and designers involves five steps, namely Analysis, Design, Development, Implementation and Evaluation. The evaluation is a multidimensional and essential step of the MOOC creation process as it is involved by all the other steps in this process. In this context, the proposed aspect-based sentiment analysis-ABSA framework can take an important role in the process of creating effective MOOCs, as shown in Figure 1. This is done by helping MOOC instructors and designers to identify and address all aspects that affect every step in the process in order to improve their separate effectiveness and the effectiveness of the entire course creation process as an organic whole.
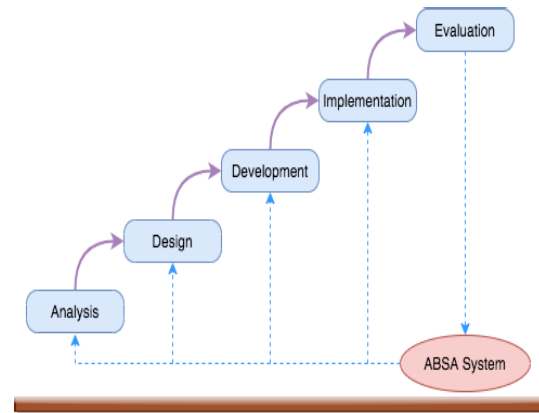

FIGURE 1. The application of the aspect-based sentiment analysis-ABSA framework in the broad process of creating effective MOOCs.

The rest of the paper is organized as follows. Section II gives a brief overview of the purpose and main objectives of this research work followed by related work presented in Section III. Main contributions are given in Section IV while MOOC-related aspects are defined in Section V. In Section VI, the architecture of the proposed weakly supervised ABSA framework is described. Experimental procedures are presented in Section VII followed by the obtained results and their analysis given in Section VIII. The conclusion and some future directions are outlined in Section IX.

II. PURPOSE AND OBJECTIVES

Our research work was driven by a concrete need to help educators and instructional designers of online courses, especially MOOCs, to identify the main aspects/factors that affect the teaching-learning process by making use of students' feedback. Therefore, the main motivation behind our research is to investigate practical solutions to gain insights into students' thought and opinion expressed in reviews, as they offer a student perspective on learning in general, and on MOOC education approach, in particular. More specifically, the main purpose of this study is to develop a weakly supervised aspect-based sentiment analysis framework capable of predicting core aspects of teaching effectiveness as they apply to a MOOC setting and assessing students' attitudes toward these aspects. In light of the purpose of this research work, we specifically formulated the following three objectives:

> **Objective 1:** *Identify and define the main aspects/ dimensions which have a crucial role in determining the effectiveness of a MOOC.*
> The core of this objective is to identify and define core aspects that influence teaching effectiveness of a MOOC, including aspects dealing with broader course-related teaching effectiveness dimensions like course, instructor, assessment, as well as more specific ones, e.g., course content, instructor skills. Moreover, a list of most relevant terms of each aspect category will be extracted.

Objective 2: *Develop a weakly supervised framework for the aspect category identification and the aspect sentiment classification.*

The main focus of this objective is to develop a weakly supervised framework that is able to identify the aspect categories discussed in a student review given the aspects defined in objective 1 and to assign a polarity to the aspect categories commented on the review based on the sentiment that is expressed in the review about it.

Objective 3: *Conduct a thorough performance evaluation of weakly supervised framework for ABSA on two real-world datasets.*

This objective entails performing a comprehensive performance evaluation of proposed ABSA framework for both prediction of the aspect categories and sentiment orientation classification using a MOOC dataset consisting of about 105k students' reviews collected from Coursera and a dataset comprising of 5989 students' feedback in traditional course settings. Additionally, performance of two deep networks, CNN and LSTM, initialized with pre-trained general-purpose embeddings and domain word embeddings will be investigated.

III. RELATED WORK

Sentiment analysis is a task of identifying and extracting users' opinion about a given topic or a product into a positive, negative, and neutral polarity, score (ranking, aggregated), or star ratings. It can be performed at a document level [10], [11], sentence level [12], [13], topic level and aspect (feature) level [14], [15]. It can further be categorized based upon the techniques used, such as, lexicon-based [16]–[18], features-based [10], [19]–[21], those using conventional machine learning approaches, i.e., Naive Bayes (NB), SVM [18], and unsupervised methods [14], and more recently deep learning-based sentiment analysis [12], [22]. A detailed description of techniques and approaches used to perform sentiment analysis is explored in the survey conducted in [23].

Sentiment analysis has been widely performed on product reviews [24], [25], sale forecasting [26], twitter data [27], [28], sarcasm detection [22], and financial domain [29] among others. For instance, multi-feature fusion sentiment classification based on deep learning is proposed in [30] to classify user's opinion expressed in movie reviews. Their method incorporates various features including statistical, linguistic knowledge, word embedding, and sentiments to create a vector representation for each sentence. The authors in [22] stated that sarcasm detection and sentiment classification are correlated for which they proposed a multitask learning-based framework using a deep neural network. The dataset was limited to less than 800 samples of one to two sentences. Ensemble of deep learning methods and traditional feature-based methods called 3W-CNN is proposed in [12]. A confidence divider component is proposed to classify the quality of CNN outcomes, which are further subjected to NB-SVM to reclassify predictions with weak confidence.

Sentiment analysis has recently gained popularity in education domain. A document level polarity assessment on students' feedback to evaluate teachers is presented in [31] using lexicon based approach. The authors employed Pearson's correlation to show the relationship between sentiment score and numerical response ratings of teacher evaluation. They further suggested that to gain additional insight into teacher evaluation aspects, aspect-based analysis is required on a large scale. The authors in [18] proposed a hybrid approach combining machine learning and lexicon-based features for sentiment analysis. They employed $tf * idf$ and lexicon features via MPQA[1] subjectivity on students' textual feedback concerning a course. The hybrid model was trained using random forest and SVM, that proved better than n-gram and bi-gram features. The dataset was limited to 1230 sentences. Rajput *et al.* [16] presented a sentiment analysis metric that highly correlates with the aggregated Likert scale scores. They utilized lexicon-based approach to evaluate teacher's performance from textual feedback and presented the results as a tag cloud and sentiment score. They used the sentiment dictionary (MPQA corpus) containing 8221 words along with their polarity. The dataset comprised of 1748 student's feedback from 63 courses collected over a period of five years.

Very recently, researchers reported the effectiveness of utilizing word embedding on online courses reviews [32]. They showed that embeddings trained on online courses having smaller specific resources are more effective than those on larger general-purpose resources. The researchers in [33] evaluated 249 MOOC courses from Class Central comprising 6393 students and their perception (satisfaction) regarding the course. They concluded that the instructor, assessment, schedule, and content are important aspects in predicting students' satisfaction whereas the course duration, major, workload, and difficulty do not contribute much in students' perception of satisfaction.

A. ASPECT-BASED SENTIMENT ANALYSIS

Aspect-based sentiment analysis is a sub-task of sentiment analysis which provides deeper understanding of the task at hand. For example, *'The course is outdated, but the teacher is good.'*, for course aspect, the polarity is negative while for teacher it is positive. Wang *et al.* [34] stated that ''*the polarity of the sentence is highly depended on both content and aspect*.''. To assign polarity to a text, it is therefore important to understand the context first by identifying the aspect.

Recently, researcher started employing various approaches to perform aspect-level sentiment classification. In [35], the authors used parameterized gates and filters to incorporate aspect information into CNN. They evaluated their model on SemEval 2014. A coarse-to-fine task transfer approach for aspect-level sentiment classification is presented in [36]. The authors used knowledge acquired from a rich-resource domain having coarse-grained aspect cate-

---

[1]http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/

gory (from Yelp recommendation dataset) to low-resource domain having fine-grained aspect terms (on SemEval'14 and twitter). They proposed a multi-granularity alignment network to address the inconsistency and mismatch between domains. The authors in [14] proposed unsupervised open information extraction strategy to monitor real-time reviews. They tested their method on the SemEval dataset. Their proposed method relied on the Stanford CoreNLP Library to utilize three linguistic resources, i.e., sentiment lexicon, WordNet and Stanford coreNLP itself. For the sentiment lexicon they aggregated the polarity values coming from SenticNet, General Inquirer vocabulary,[2] and the the MPQA dictionary.

However, there is a limited research on aspect-based sentiment analysis in the education domain. An earlier work reported on document sentiment analysis which was further explored for faculty performance evaluation from students' comments as feedback [10]. The overall document level faculty performance score was calculated from individual scores on presentation, communication, knowledge, and presentation of the faculty members. The dataset comprised of around 5000 positive and negative comments. Naive Bayes and SVM classifiers were employed with an accuracy of 72.80% and 81% respectively. In [20], the researchers proposed an aspect based sentiment analysis tool to evaluate higher education institution's reputation from users' reviews on Facebook and Twitter. The tool used the Stanford coreNLP library to perform sentiment analysis and they reported 72.56% accuracy. The authors in [37] proposed domain oriented aspect level sentiment analysis on students' feedback using OpenNLP parser for POS tagging and sentiWordNet lexical resources for defining the wordScore. Domain specific ontology on their own dataset is developed to detect aspects.

The closest research we could find as of today is presented by Sindhu *et al.* in [21]. They proposed a two-layered LSTM model for aspect-based opinion mining of students' feedback. The first layers was used to predict aspects from the feedback while the second finds the polarity, i.e., negative, neutral, and positive. They tested the system on their own data set consisting of nearly 5990 students' feedback of the past five years. Six aspects were manually tagged including pedagogy, teachers' knowledge, experience, assessment, behaviour, and general. They achieved a 91% accuracy on aspect extraction and 93% on polarity detection. Similar work was presented in [38], where the authors conducted analysis of students' feedback to improve the teaching-learning process. The aspects are extracted using Stanford NLP parser which are further pruned to determine most frequent aspects. Teaching and course evaluation aspects are extracted and processed for sentiment aggregation on only 1000 students' comments from social media. They achieved 80% and 81% F1 score respectively on aspect extraction and polarity assessment with 72% overall sentiment score.

---

[2]http://www.wjh.harvard.edu/inquirer/spreadsheet_guide.htm

B. WEAKLY SUPERVISED LEARNING

One issue in the education domain is the unavailability of the labelled data. Khan *et al.* [39] stated that ''*the core hindrance in the application of supervised algorithms or domain specific sentiment lexicons is the unavailability of sentiment labeled training datasets for every domain*''. The data that are labelled manually by the institutes are either too small or private. To train deeper neural networks on feedback obtained on massive open online courses (MOOCs) from a large number of students, we adopted a weakly supervised learning approach explained in section VI-D. Such an approach has been used in other domains. For example, the authors in [40] identified the main words that contributed towards the sentiment classification. They identified those words which contribute towards discriminating positive and negative sentences via a weakly supervised CNN method. They implemented a word attention mechanism for identifying high-contributing words. The words are obtained as a result of classification via class activation map [41] by using weights from the last fully connected layer of the learned CNN model. Word2Vec, GloVe, and FastText methods are used to perform low-level embedding. Vectors obtained from these embedding are concatenated before feeding to the CNN. The English (IMDB dataset) and Korean (WATHCA) movie reviews are evaluated.

The authors in [17] proposed sentiment lexicons augmentation utilizing weakly supervised technique and direct propagation of sentiment signals to estimate reputation popularity of tweets. They also proposed a polar fact filter for differentiating between reputation-bearing and reputation neutral tweets. They used the RapLab2013 dataset containing tweets in English and Spanish. This dataset is limited to banking, music, universities, and automotive domain. The paper indicates that reputation polarity can be effectively predicted using pointwise mutual information (PMI).

Our proposed approach is similarly concerned with examining opinions of students expressed in textual reviews, but it differs from previous studies in three directions. First, unlike aforementioned approaches that investigated handling of students' opinion mainly from small-scale feedback in traditional course delivery setting, we investigated fine-grained opinion polarity toward course-related aspects that may affect teaching effectiveness by analyzing large-scale students' feedback in both course delivery modes, online mode-MOOC and face-to-face delivery mode. More specifically, these aspects include broader-level dimensions like course, instructor, technology, and specific-level ones such as course content, course structure, instructor skills, etc. Second, our approach takes advantage of the strengths of the weakly supervised technique for aspect-level sentiment analysis. This technique eliminates the need for most, if not all, human annotated data which are either not readily available or very expensive to obtain. Moreover, to the best of our knowledge our study is the first that explores aspect-based sentiment analysis employing weakly supervised approach in education domain on a large scale-such as MOOCs. Third,

two deep neural networks, namely CNN and LSTM, using both pre-trained general-purpose embeddings and domain-specific word embeddings are used to assess students' sentiment polarities expressed in reviews.

IV. CONTRIBUTIONS

Most work reported in the education domain is carried out at a document level without taking into consideration different aspects (features). Our paper differs from existing work in the following aspects:

1) We present a framework for aspect-based sentiment analysis on two real-world datasets.
2) In addition to high-level aspects identification, we examined the polarity of fine-grained aspects in the education domain.
3) We propose a weakly supervised learning approach to handle a large amount of students' feedback.
4) The proposed architecture can easily be used to evaluate students' feedback not only to assess the pedagogy and course related aspects, but teacher's knowledge and performance in a traditional brick-and-mortar or MOOC setting.

V. ASPECTS DEFINITION

Online courses, especially MOOCs, are multidimensional constructs where teaching effectiveness depends on various factors which affect that in different ways. These factors primarily include dimensions dealing with the course design, execution, evaluation, factors concerning interpersonal relationships and those related to the technology for delivering the content [8]. In line with these findings, we identified a set of aspects that are crucial on determining the effectiveness of MOOCs. These aspects, categorized into two groups: specific MOOC-related aspects and broader MOOC-related aspects, are defined in the following.

Specific MOOC-related aspects:

- **content** is devoted to the subject matter of the MOOC including learning material, projects, and learning examples.
- **structure** refers to the structural dimensions of the MOOC, such as structure of modules, appropriateness of instructional approach along with clearness of learning objectives.
- **knowledge** encompasses both theoretical and practical understanding that an instructor has for the subject matter.
- **skill** refers to the teaching pedagogy of the instructor including appropriateness of teaching strategies.
- **experience** is instructor's experience gained over time including teaching, mentoring and supervision.
- **assessment** deals with evaluation tools or methods used by instructor or peers (classmates) to assess knowledge and skills gathered by students.
- **technology** refers to the information technology used for delivering of the learning material, e.g., quality of audio and video lectures.

- **interaction** indicates interactions of students in a classroom including student-to-student interactions, e.g, the extent to which the course encouraged students to learn from their peers, and student-to-instructor interactions like responsiveness and accessibility of the instructor.
- **general** concerns students' comments covering the entire MOOC rather than any particular aspect of it.

Broader MOOC-related aspects are higher-level drivers that have very important role on the effectiveness of MOOCs. These four aspects are: 1) the course that covers both content and structure aspect, 2) the instructor that includes knowledge, skill and experience of the instructor, and 3) the assessment, and 4) the technology. For aspects 3 and 4 apply the same definitions as in specific MOOC-related aspects.

VI. ARCHITECTURE OF THE PROPOSED FRAMEWORK

The architecture of the proposed weakly supervised framework for aspect-based sentiment analysis comprises 4 main components, namely user input information, aspect category learning, weak label propagation, and aspect category polarity. Figure 2 illustrates the high-level conceptual architecture of the ABSA framework.

A. USER INPUT INFORMATION

This module entails the user-provided seed information that serves as a supervision signal for the aspect category learning task. The module can handle two main types of seed information including manually annotated reviews and aspect-related terms. The former represents a small number of manually tagged reviews provided by a user for each aspect category while the latter consists in providing a small set of keywords that are related to a given aspect category.

Supervision signal is formally defined as follows. Let $R = \{r_1, r_2, \ldots, r_l, r_{l+1}, r_{l+2}, \ldots, r_n\}$ be a collection of n reviews, $A = \{w_1, w_2, \ldots, w_k\}$ a set of k keywords in an aspect category, and let $C = \{c_1, c_2, \ldots, c_n\}$ be a finite set of labels for n aspect categories.

Then manually annotated reviews denoted by R_L are the first l reviews $r_i, i \in L = \{1, .., l\}$ labeled according to $c_i \in C$, while the remaining $t = n - l$ reviews $r_i, i \in T = \{l+1, .., n\}$ denoted by R_U are unlabeled reviews. Aspect-related keywords denoted by A_K is the tuple $A_K = \{w | w \in A^*, A^* \subset A\}$.

B. ASPECT CATEGORY LEARNING

Manually annotated reviews R_L provided from user input information module undergo a preprocessing step where a set of relevant terms is extracted using term frequency-inverse document frequency weighting-*tf* $*idf$ technique. The extracted terms are the representative terms of each aspect category c_i and thus will receive the same treatment as the aspect category-related terms A_K.

After obtaining a set of relevant terms that are correlated with each aspect category, the word2vec's skip-gram model [42] is used to learn and generate their d-dimensional
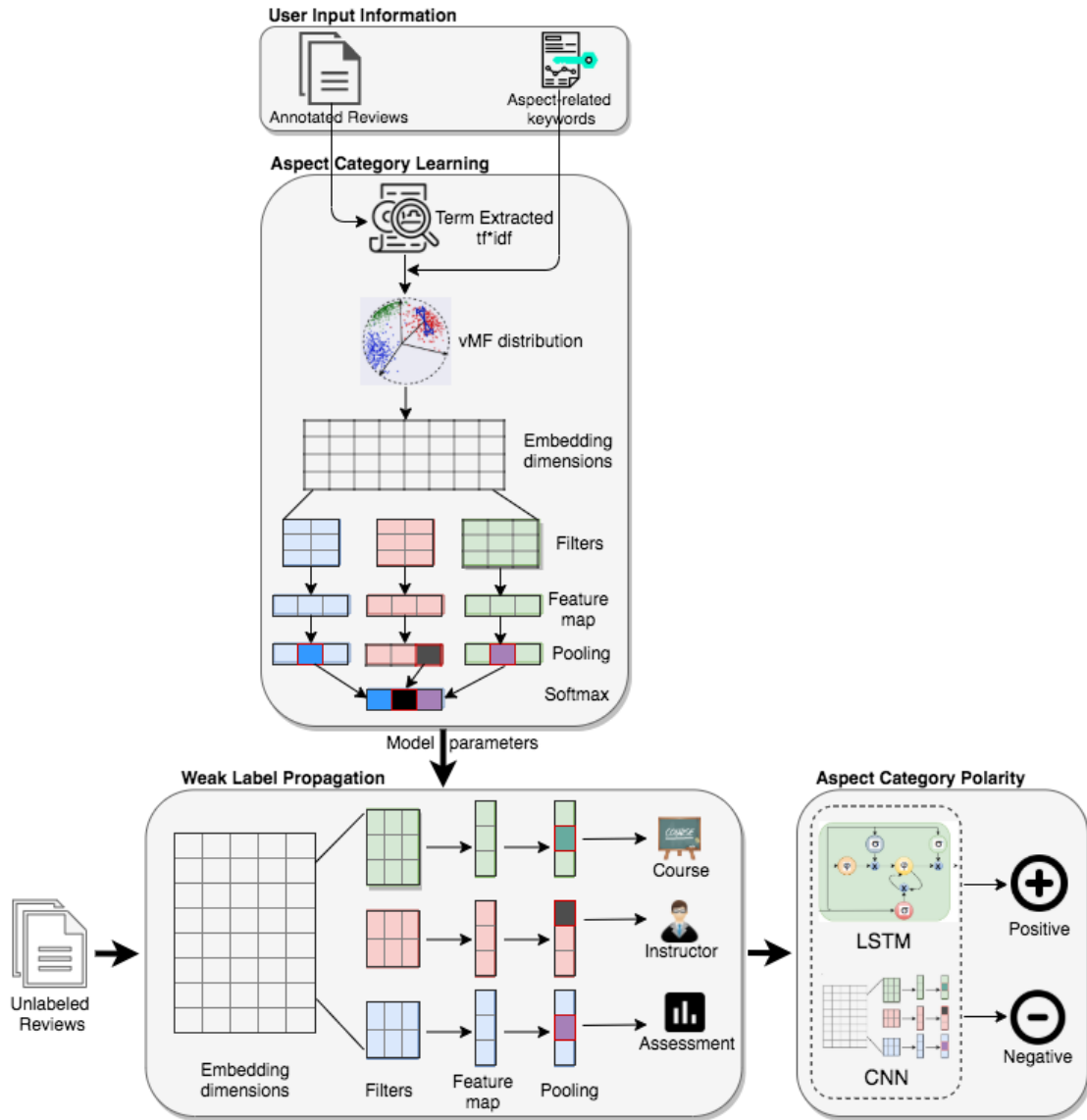
FIGURE 2. High-level conceptual architecture of the proposed weakly supervised framework for aspect-based sentiment analysis.

embeddings. Word embeddings will be used to extract the surrounding context of the aspect-related terms by retrieving their top-t closest terms residing within a unit hypersphere in $\mathbb{R}^d$. A unit hypersphere represents the distributional semantic space of a given aspect category and it is modeled using von Mises-Fisher (vMF) distribution [43] defined in Equation 1.

$$f_d(x|\mu, \kappa) = c_d(\kappa)e^{\kappa\mu^T x} \qquad (1)$$

where, μ is the mean direction with a norm equal to one ($||\mu|| = 1$), κ denotes the concentration parameter with positive values ($\kappa \geq 0$), d represents embedding dimensions ($d \geq 2$) and $c_d(\kappa)$ is the normalization constant which is defined in Equation 2.

$$c_d(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}I_{p/2-1}(\kappa)} \qquad (2)$$

where $I_s(\kappa)$ is the modified Bessel function of the first kind at order s.

The semantic information on a unit hypersphere is characterized by the mean direction μ by generating relevant word embeddings around it whose concentration is controlled by parameter κ [44]. The maximum likelihood estimates for the parameters μ and κ of the vMF distribution $f_d(x|\mu, \kappa)$ are computed with respect to the seed terms correlated to aspect categories provided by the user.

Seed information expanded with top-t closest terms extracted by vMF distribution model are incorporated to the embedding layer of deep learning CNN model. The information received from embedding layer will go through a convolution layer with filters of different size to generate the feature map. Then a maxpooling operation is performed on

the feature map to obtain features with the maximum values. Finally, a *softmax* activation function generates a probability distribution over all possible aspect categories. The aspect category with the highest probability value is chosen as a predicted class for the corresponding input review.

C. WEAK LABEL PROPAGATION

The aspect category learning module generates an initial deep learning CNN model for prediction of aspect categories c_i using labeled reviews R_L. The generated model is used to label all unlabeled students' reviews R_U collected from MOOCs. Training the deep network model with only labeled reviews can not achieve the best performance as it does not consider the information encoded in the unlabeled reviews. To address this issue, the study in [44] proposed a widely used technique for semi-supervised learning known as the self-training [45]. The idea behind the self-training technique is to improve the performance of deep network model by repeatedly refining its parameters by using the training set containing the most confident unlabeled reviews and their predicted labels. In other words, the model teaches itself using its own predictions and hence the learning strategy is also referred to as self-teaching or bootstrapping [45].

The deep learning model is the same CNN architecture as presented in aspect category learning module including the following layers: an embedding layer followed by the convolutional layer with different filters where each filter will be slid across the embedding matrix to generate a 1D feature map. Next, a maxpooling layer is used to downsample the input feature map. Finally, the *softmax* layer depicts a probability distribution over broader/specific course-related aspect labels.

D. ASPECT CATEGORY POLARITY

This module is responsible for the determination of polarity, i.e., positive (P) and negative (N), of each aspect category as produced in weak label propagation module. We specifically placed the focus on the negative and the positive constructive feedback omitting the neutral ones. The former focuses on behavior or a task that was not perform successfully thus indicating that should be fixed and not repeated. The later focuses on the behavior or the task that went well as expected or desired and that should be continued. We omitted neutral feedback as they do not play any role in the learning-teaching process with respect to course delivery improvement and efficiency.

To perform the aspect polarity assessment task, two deep networks including CNN and LSTM will be used. The networks take both reviews R and aspect categories C as an input denoted by D and depict a vector of class probability scores denoted by $f_\theta : D \times C \rightarrow \{P, N\}$, where θ indicates the network parameters. Two other major steps occur in-between of input space and class probability scores. The first step is feature extraction, $\Psi_\theta : D \rightarrow \mathbb{R}^d$, where the input information is mapped to a d-dimensional feature vector. The second step comprises a fully connected layer used on top of Ψ_θ and

followed by a *softmax* layer depicting aspect polarity scores. The class polarity (P or N) with the highest probability score will be assigned to the given input information D.

VII. EXPERIMENTS

This section provides details about the deep network models used for both the prediction of aspect categories and aspect sentiment classification as well as of the dataset used for the experiments.

A. DEEP NETWORK MODELS

The deep network applied to learn the aspect categories and label propagation is a CNN architecture composed of several types of layers as depicted in Figure 3. The input layer consists of input feature vectors built of MOOC reviews followed by an embedding layer containing embeddings of size $100 \times 100d$. The output of the embedding layer is fed into a convolutional layer comprising 20 1D convolution filters of various sizes s, for $s = \{2, 3, 4, 5\}$ and *ReLU* activation function. Next, a 1D global maximum pooling operation is applied on top of each convolution blocks to compute the maximum value for each of the blocks. Then, a deep concatenation layer is used to concatenate the result of convolution blocks along the depth dimension into a single vector creating the input of the next fully-connected layer with *ReLU*. Finally, a dense layer with *softmax* activation function is used to compute probability distribution of each aspect category.
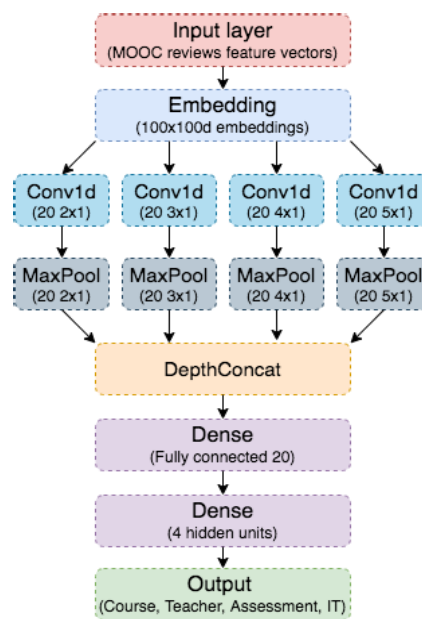

FIGURE 3. CNN Architecture for prediction of the aspect categories.

To identify the opinion orientation of students toward aspect categories commented on the reviews, we used two deep network models as illustrated in Figure 4. The first model is a simple CNN architecture (Figure 4(a)) composed of an input, an output and 5 hidden layers. The first input layer comprises reviews along with aspect categories followed by
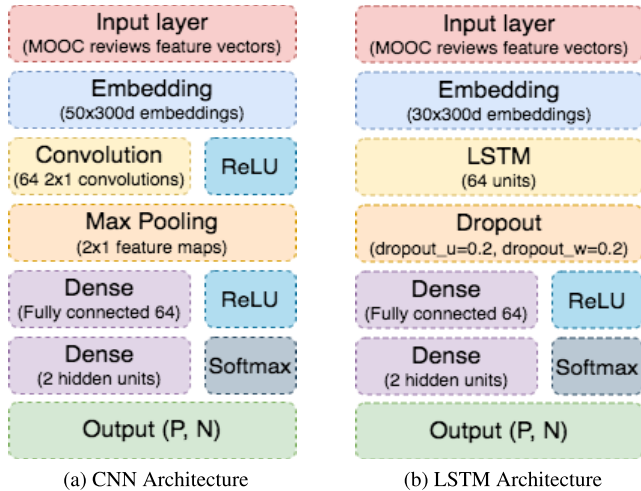
(a) CNN Architecture (b) LSTM Architecture

FIGURE 4. Deep network models for prediction of the aspect category polarity.

an embedding layer with word embeddings of size $30 \times 300d$. Next comes a convolution layer with 64 1D convolution filters of size 2 and a *ReLU* activation function is applied. Finally, two fully-connected layers are used on top of the maxpooling layer. The first dense layer is composed of 64 units and *ReLU* while the second dense layer contains 2 units and a *softmax* function for outputting probability distribution over two sentiment orientation denoted by P and N.

The second model used to identify the students' opinion orientation with respect to aspect categories is a deep network whose architecture is similar to that of CNN with a slight difference in configuration as shown in Figure 4(b). Instead of using the convolution and maxpooling layers, it is the LSTM and dropout layers that are applied to this architecture.

B. DATASET
Elia *et al.* [46] assessed learners' satisfaction in collaborative online courses through a big data approach. They suggested that large-scale data may support learning managers to enhance decision making processes. We, therefore, conducted the experiments on a large-scale education dataset containing students' reviews on MOOCs collected from Coursera. All reviews are in English language. Table 1 provides some dataset statistics including the number of reviews and their lengths in terms of sentences and words.

The dataset contains students' reviews having comments about various aspects related to a given MOOC. There are some reviews in which only one single aspect is commented on. For example, the review "*The teacher is one of the best*

TABLE 1. Dataset statistics.

Data	Value
# of reviews	104999
max length	421 sentences / 350 words
avg length	1.87 sentences / 8.36 words
min length	1 sentence / 1 word
Review length std	18.32 words

i have ever seen, truly an expert in this field of science.", is just about the instructor aspect. There exists some other reviews comprising of comments about almost all aspects. For instance, in the given review "*Fantastic course i have learned so much. Great videos full of information and great instructors*", the student has commented on the following aspects course, instructor, and technology.

VIII. RESULTS AND ANALYSIS
In this section we provide a set of results obtained from experiments conducted to investigate the performance of proposed ABSA framework on accomplishing two major tasks, the aspect category identification and the aspect sentiment classification. The first part entails results achieved by the proposed system on performing the first task.

In the initial stage of the experiment, we analyzed the expansion of seed information that is used as a supervised data to feed the model. Table 2 depicts the list of aspect-related keywords provided as a supervision signal expanded with top-10 closest terms located within the same unit hypersphere and extracted using vMF distribution model. We provided 3 keywords that are associated with and can describe each aspect category. Specifically, keywords such as *course*, *content*, *structure* are provided from us for the course aspect, *instructor*, *teacher*, *teaching* for instructor, *assessment*, *assess*, *grading* for assessment and keywords like *technology*, *audio*, *video* for technology aspect category.

We observe from Table 2 that all extracted terms are semantically related to the terms provided by us as a supervision signal. This highlights the usefulness of the unit hypersphere on projecting aspect terms along with their surroundings into a compact semantic space.

Table 3 shows top-10 terms extracted from the supervision signal coming from manually annotated reviews. Specifically, we chose to present the relevant terms generated using $tf * idf$ on 30 manually labeled reviews provided for each aspect category as they gave the highest prediction accuracy (Figure 5).

We investigated how sensitive the proposed prediction model is to the amount of supervision signal coming from labeled reviews. Various number of labeled reviews are provided for each aspect category, starting from 10 to 100 with an interval of 10. The performance obtained from the proposed model on predicting aspect categories including those dealing with general-level course-related dimensions denoted by GL, as well as more specific ones denoted by FG, is shown in Figure 5. Specifically, we present the performance with respect to macro- and micro-F1 score as they use different approaches to evaluate the overall performance of the model. Macro-F1 score is a simple arithmetic mean of F1 scores per each aspect category while Micro-F1 assesses the performance of the model over all instances appearing together in the aspect categories.

It is apparent from Figure 5 that the highest performance is achieved, in both cases, when the number of labeled reviews is 30. Specifically, an F1 score of 80.64% and 65.64% is

TABLE 2. Top-10 closest terms associated with each aspect category.

Aspect category	Top-10 closest terms
course	content, pacing, pedagogy, syllabus, structure, layout, progression, coursework, material, didactics
instructor	teacher, instructor, lecturer, presenter, tutor teaching, professor, lecturing, calm, narration
assessment	marking, assessment, grading, rubric, criteria, rubrics, automatic, peers', checks, submissions
technology	audio, video, pauses, placing, mismatch, professionalism, biases, factor, powerpoints, inconsistencies

TABLE 3. Top-10 terms extracted from labeled reviews for each aspect using $tf * idf$.

| Aspect category	Top-10 extracted terms
course	objective, great, structured, content, learned, thank, interesting, really, useful, liked
instructor	knowledge, subject, good, teacher, sir, excellent, regarding, prepared, finance, informative
assessment	grading, great, good, really, lot, pace, took, content, assignment, overall
technology	audio, great, good, instructor, quality, overall, practical, prepared, fantastic, video
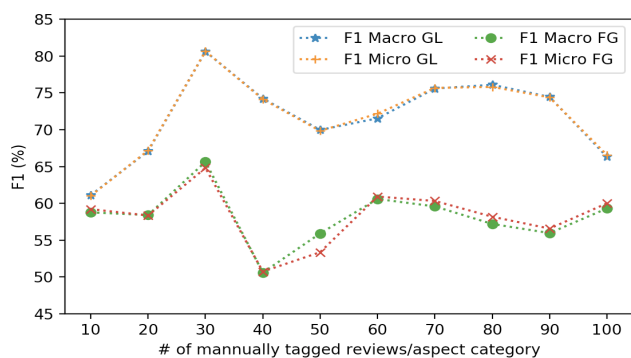
FIGURE 5. Performance of the model with respect to the number of manually annotated reviews per aspect category.

FIGURE 6. Performance of model with respect to the number of keywords associated to each aspect category.

achieved for the broader course-related aspect and the specific course-related aspect respectively. This may have happened because the terms extracted using $tf * idf$ weighting scheme (Table 3) have shown to have high discriminative power necessary for the prediction task. From the graph we can also note that there is gap in F1 score achieved by the model on broader course related aspects and more specific ones. A possible reason for this is that specific-level course-related aspects are not clearly distinguished from each other and are overlapped sometimes with more than one other aspect.

Next, we extended the experiment to examine the effect of seed keywords associated to aspect categories would have on the performance of the model. To accomplish this task, we tuned the number of keywords provided as a supervision signal and compared the model performance. We started with 1 to 10 keywords with a step size of 2. It should be noted that the first keywords given for each aspect category represent their corresponding names. Figure 6 clearly shows that the best performance is obtained from the two groups of aspects when 3 aspect-related keywords are provided as input, achieving an F1 score of 75.78% for general-level aspects and 63.10% for specific-level aspects respectively. In fact, providing only names of broader aspect categories yields the best performance with an F1 score of 86.13%. This could be attributed to the compactness of the semantic space
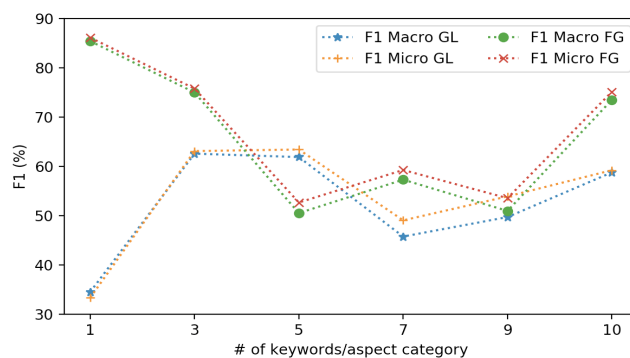
modeled by a unit hypersphere in which all terms including seed keyword and its top-10 closest terms have zero degree of overlap.

In the next set of experiments, we examined the performance of proposed model on performing the second task in ABSA, that is, the aspect category sentiment classification. To achieve this, we conducted experiments using two deep networks LSTM and CNN whose architectures are shown in Figure 4. Word embeddings generated from three different types of static pre-trained general-purpose models, namely FastText [47], GloVe [48], Word2Vec [42], and the domain-specific MOOC word embeddings are used for experimenting. The choice of the deep learning architectures and pre-trained word embedding models for aspect sentiment classification is based on our thorough analysis of semantically rich representations for classification task conducted in [49]. The obtained results with respect to precision, recall and F1 score are given in Table 4 and Table 5.

TABLE 4. Performance of CNN using different word embeddings.

Model	Precision		Recall		F1 score	
	macro	weight	macro	weight	macro	weight
FastText	82.87	82.61	81.36	82.31	81.75	82.10
GloVe	81.69	81.73	81.21	81.74	81.39	81.67
Word2Vec	80.05	80.48	80.33	80.27	80.13	80.31
MOOC	83.14	82.80	81.27	82.33	81.70	82.08

TABLE 5. Performance of LSTM using different word embeddings.

Model	Precision		Recall		F1 score	
	macro	weight	macro	weight	macro	weight
FastText	81.59	81.49	80.63	81.40	80.92	81.26
GloVe	81.99	81.98	81.38	81.98	81.59	81.89
Word2Vec	81.95	81.84	80.99	81.74	81.28	81.61
MOOC	81.37	81.32	80.59	81.28	80.84	81.16

As shown in Table 4 and Table 5, CNN performs slightly better than LSTM among all types of embeddings. A reasonable explanation for this is that LSTM could not be able to capture long-range semantic dependencies to deriving sentiment polarity as our MOOC dataset contains students' reviews comprising of very short sentences. On the contrary, CNN is able to extract position-invariant features [50], which are keywords that determine sentiment polarity and therefore it achieves better performance.

We observe from Table 4 that a slight performance improvement of CNN is achieved using FastText and MOOC domain word embeddings. FastText relies on character level information and it is capable of capturing the semantic information about suffixes/prefixes and short words that may appear as n-grams of other words. On the other side, MOOC embeddings are domain-specific word embeddings [7], which allow to capture semantic relations and technical vocabulary specifically related to the domain.

A. COMPARISON WITH THE BASELINE APPROACH

We compared our approach to the one presented by Sindhu *et al.* in [21] on their dataset. We specifically selected this dataset as a benchmark and their proposed approach to compare with ours, as this is the only manually labeled dataset available as of today. This dataset comprises of 5989 student's feedback sentences collected over a period of five years, which are categorized into six aspect categories. These include pedagogy, knowledge, experience, assessment, behaviour, and general. The data was manually labelled into positive, negative, and neutral polarity. The number of instances per aspect category are depicted in Table 6.

TABLE 6. Number of instances per aspect category.

No.	Aspect Category	# of instances
1	Assessment	139
2	Behaviour	844
3	Experience	297
4	Teaching Skills	1656
5	Knowledge	1186
6	General	1867
7	Total	5989

Students' feedback vary widely in length, ranging from 1 to 86 words, with an average length of 6.43 words (std: 4.76).

Our weakly supervised approach is applied to label the students' feedback into positive, negative, and neutral polarity. Weak supervision signal constituted by 3 seed keywords for each sentiment orientation is used as an input to train the
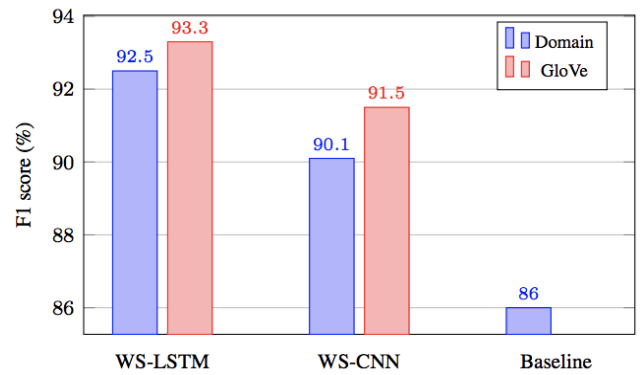
FIGURE 7. Performance of our approach vs baseline approach on sentiment classification.

model. Specifically, keywords *bad*, *dissappointed*, *biased* are used as supervised data for positive sentiment, keywords *excellent*, *great*, *awesome* for negative sentiment and *comment*, *teacher*, *course* are the neutral sentiment-related keywords.

An automatically annotated dataset is used to train the deep network model for the aspect sentiment classification as presented in paper [21]. The architecture of the network is relatively simple, comprising an embedding layer, dropout layer followed by a 196-unit LSTM, and the output layer depicting a probability distribution over three sentiment orientations, i.e., positive, neutral, and negative. In addition, we tested a CNN deep network which had a very similar architecture as LSTM network but with a CNN layer instead of LSTM. GloVe pre-trained word vectors and domain word embeddings with 100d are used for evaluation of the models.

It can be seen from the graph that both deep networks, LSTM and CNN, trained on the dataset labeled using our weakly supervised approach, denoted as WS-LSTM and WS-CNN, achieved better performance than the baseline approach with respect to aspect sentiment classification. More precisely, the best performance is obtained by LSTM with GloVe word embeddings achieving an F1 score of 93.30%. This is in contrast to the results with our MOOC dataset when CNN performed better. This can be explained by the fact that the baseline dataset comprises feedback of sentence length with no sentential connectives ('but', 'and') in which only a single teacher-related aspect is commented on. Moreover, a short sentence feedback allows the LSTM to capture semantic dependencies over the whole sentence to derive its sentiment.

IX. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework for aspect-based sentiment analysis of students' feedback of MOOCs. The framework took advantage of the use of weak supervision strategy for prediction of the aspect categories that are critical factors in determining the effectiveness of online courses in general. Next, the framework analysed the attitude of students

toward these aspects expressed on given comments. Two groups of aspects are identified. The first group is about wider MOOC-related aspects, whereas the second one is concerned with more specific aspects. Two supervised data types are fed into a deep network in the framework for ABSA. One type contained a set of keywords associated to each aspect category and the other type comprised of a small amount of manually annotated reviews. The proposed framework is tested and validated on two real-life datasets containing students' feedback gathered from Coursera and students' feedback in traditional course settings. The obtained performance, F1 score of 86.13% for aspect category identification (broader MOOC-related aspects) and 82.10% for aspect sentiment classification (CNN+FastText) demonstrated that the proposed framework is reliable and comprehensive. Moreover, it outperformed the baseline model on a manually labelled dataset, achieving 93.3% F1 score.

The framework proposed in this paper is of great practical relevance to educators and instructional designers as it can help them to gain insights and identify the most important aspects of the course based on students' feedback. These insights can then be used to adopt the course design and pedagogical model in order to improve the course delivery and increase its efficiency. This framework is particularly applicable in MOOCs-like courses where the drop-out rates are usually very high. The importance and influence of the course design and pedagogical model on the retention rates and overall performance of the students has been advocated by Xing and Wanli (2019) [51].

The framework is applied on education domain but it can be easily used for other domains with a slight modification of input parameters, i.e., supervised data signal, thus future work will concentrate on generalizability of the proposed framework by testing it on other non-education domains especially on benchmark dataset. Another promising direction of future work is to examine contextualized word embeddings models like BERT or ELMo as they can handle both out-of-vocabulary and disambiguation issue.

ACKNOWLEDGMENT

REFERENCES

[1] K. Masters, "A brief guide to understanding moocs," *Internet J. Med. Edu.*, vol. 1, no. 2, pp. 1–2, 2011.

[2] S. L. Schneider and M. L. Council, "Distance learning in the era of COVID-19," *Arch. Dermatological Res.*, vol. 8, pp. 1–2, May 2020.

[3] H. Khalil and M. Ebner, "MOOCs completion rates and possible methods to improve retention—A literature review," in *EdMedia + Innovate Learning*, J. Viteli and M. Leikomaa, Eds. Waynesville, NC, USA: Association for the Advancement of Computing in Education, 2014, pp. 1305–1313.

[4] A. Itani, L. Brisson, and S. Garlatti, "Understanding learner's drop-out in MOOCs," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.* Cham, Switzerland: Springer, 2018, pp. 233–244.

[5] A. Kurti and F. Dalipi, "Bridging the gap between academia and industry: Lessons learned from a graduate it professional development program," in *Proc. 2nd Annu. Int. Conf. Eng. Edu. Teaching*, 2017, p. 27.

[6] A. S. Imran and F. A. Cheikh, "Multimedia learning objects framework for e-learning," in *Proc. Int. Conf. E-Learn. E-Technol. Edu. (ICEEE)*, Sep. 2012, pp. 105–109.

[7] Z. Kastrati, A. S. Imran, and A. Kurti, "Integrating word embeddings and document topics with deep learning in a video classification framework," *Pattern Recognit. Lett.*, vol. 128, pp. 85–92, Dec. 2019.

[8] Z. Kastrati, A. Kurti, and J. Hagelbäck, "The effect of a flipped classroom in a SPOC: Students' perceptions and attitudes," in *Proc. 11th Int. Conf. Edu. Technol. Comput. (ICETC)*, 2019, p. 246–249.

[9] F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in *Proc. IEEE Global Eng. Edu. Conf. (EDUCON)*, Apr. 2018, pp. 1007–1014.

[10] N. D. Valakunde and M. S. Patwardhan, "Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process," in *Proc. Int. Conf. Cloud Ubiquitous Comput. Emerg. Technol.*, Nov. 2013, pp. 188–192.

[11] L. Kong, C. Li, J. Ge, F. Zhang, Y. Feng, Z. Li, and B. Luo, "Leveraging multiple features for document sentiment classification," *Inf. Sci.*, vol. 518, pp. 39–55, May 2020.

[12] Y. Zhang, Z. Zhang, D. Miao, and J. Wang, "Three-way enhanced convolutional neural networks for sentence-level sentiment classification," *Inf. Sci.*, vol. 477, pp. 55–64, Mar. 2019.

[13] X. Ma, J. Zeng, L. Peng, G. Fortino, and Y. Zhang, "Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis," *Future Gener. Comput. Syst.*, vol. 93, pp. 304–311, Apr. 2019.

[14] M. Dragoni, M. Federici, and A. Rexha, "An unsupervised aspect extraction strategy for monitoring real-time reviews stream," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 1103–1118, May 2019.

[15] R. Ma, K. Wang, T. Qiu, A. K. Sangaiah, D. Lin, and H. B. Liaqat, "Feature-based compositing memory networks for aspect-based sentiment classification in social Internet of Things," *Future Gener. Comput. Syst.*, vol. 92, pp. 879–888, Mar. 2019.

[16] Q. Rajput, S. Haider, and S. Ghani, "Lexicon-based sentiment analysis of teachers' evaluation," *Appl. Comput. Intell. Soft Comput.*, 2016, pp. 1–12.

[17] A. Giachanou, J. Gonzalo, and F. Crestani, "Propagating sentiment signals for estimating reputation polarity," *Inf. Process. Manage.*, vol. 56, no. 6, Nov. 2019, Art. no. 102079.

[18] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *Proc. Int. Conf. Res. Innov. Inf. Syst. (ICRIIS)*, Jul. 2017, pp. 1–6.

[19] N. M. Saeed, N. A. Helal, N. L. Badr, and T. F. Gharib, "An enhanced feature-based sentiment analysis approach," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 10, no. 2, p. e1347, 2019.

[20] L. Balachandran and A. Kirupananda, "Online reviews evaluation system for higher education institution: An aspect based sentiment analysis tool," in *Proc. 11th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA)*, Dec. 2017, pp. 1–7.

[21] I. Sindhu, S. M. Daudpota, K. Badar, M. Bakhtyar, J. Baber, and M. Nurunnabi, "Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation," *IEEE Access*, vol. 7, pp. 108729–108741, 2019.

[22] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, and E. Cambria, "Sentiment and sarcasm classification with multitask learning," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 38–43, May 2019.

[23] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: Comparative analysis and survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 459–483, Dec. 2016.

[24] J. Khan and B. S. Jeong, "Summarizing customer review based on product feature and opinion," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, vol. 1, Jul. 2016, pp. 158–165.

[25] C. Wan, Y. Peng, K. Xiao, X. Liu, T. Jiang, and D. Liu, "An association-constrained LDA model for joint extraction of product aspects and opinions," *Inf. Sci.*, vol. 519, pp. 243–259, May 2020.

[26] Z.-P. Fan, Y.-J. Che, and Z.-Y. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the bass model and sentiment analysis," *J. Bus. Res.*, vol. 74, pp. 90–100, May 2017.

[27] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 1–18.

[28] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Gener. Comput. Syst.*, vol. 106, pp. 92–104, May 2020.

[29] M. Atzeni, A. Dridi, and D. Reforgiato Recupero, "Using frame-based resources for sentiment analysis within the financial domain," *Prog. Artif. Intell.*, vol. 7, no. 4, pp. 273–294, Dec. 2018.

[30] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on multi-feature fusion," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1245–1259, Jul. 2019.

[31] P. Kaewyong, A. Sukprasert, N. Salim, and F. A. Phang, "The possibility of students' comments automatic interpret using lexicon based sentiment analysis to teacher evaluation," in *Proc. Int. Conf. Artif. Intell. Comput. Sci. (AICS)*, 2015, pp. 179–189.

[32] D. Dessí, M. Dragoni, G. Fenu, M. Marras, and D. R. Recupero, "Deep learning adaptation with word embeddings for sentiment analysis on online course reviews," in *Deep Learning-Based Approaches for Sentiment Analysis*. Singapore: Springer, 2020, pp. 57–83.

[33] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Comput. Edu.*, vol. 145, Feb. 2020, Art. no. 103724.

[34] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[35] B. Huang and K. Carley, "Parameterized convolutional neural networks for aspect level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1091–1096.

[36] Z. Li, Y. Wei, Y. Zhang, X. Zhang, and X. Li, "Exploiting coarse-to-fine task transfer for aspect-level sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4253–4260.

[37] N. Soe and P. T. Soe, "Domain oriented aspect detection for student feedback system," in *Proc. Int. Conf. Adv. Inf. Technol. (ICAIT)*, Nov. 2019, pp. 90–95.

[38] G. S. Chauhan, P. Agrawal, and Y. K. Meena, "Aspect-based sentiment analysis of students' feedback to improve teaching–learning process," in *Proc. Inf. Commun. Technol. Intell. Syst.* Singapore: Springer, 2019, pp. 259–266.

[39] F. H. Khan, U. Qamar, and S. Bashir, "Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio," *Artif. Intell. Rev.*, vol. 48, no. 1, pp. 113–138, Jun. 2017.

[40] G. Lee, J. Jeong, S. Seo, C. Kim, and P. Kang, "Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network," *Knowl.-Based Syst.*, vol. 152, pp. 70–82, Jul. 2018.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[43] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, 2005.

[44] Y. Meng, J. Shen, C. Zhang, and J. Han, "Weakly-supervised neural text classification," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Oct. 2018, p. 983.

[45] X. Zhu and A. Goldberg, *Introduction to Semi-Supervised Learning*. San Rafael, CA, USA: Morgan & Claypool, 2009.

[46] G. Elia, G. Solazzo, G. Lorenzo, and G. Passiante, "Assessing learners' satisfaction in collaborative online courses through a big data approach," *Comput. Hum. Behav.*, vol. 92, pp. 589–599, Mar. 2019.

[47] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 52–55.

[48] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, vol. 14, 2014, pp. 1532–1543.

[49] Z. Kastrati, A. S. Imran, and S. Y. Yayilgan, "The impact of deep learning on document classification using semantically rich representations," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1618–1632, Sep. 2019.

[50] B. Wang, "Disconnected recurrent neual networks for text categorization," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2311–2320.

[51] W. Xing, "Exploring the influences of MOOC design features on student performance and persistence," *Distance Edu.*, vol. 40, no. 1, pp. 98–113, Jan. 2019.

ZENUN KASTRATI received the master's degree in computer science through the EU TEMPUS Programme developed and implemented jointly by the University of Pristina, Kosovo, the Universite de La Rochelle, France, and the Institute of Technology Carlow, Ireland, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2018. He is currently a Postdoctoral Researcher with the Department of Computer Science and Media Technology, Linnaeus University, Sweden. He is the author of several articles published in international journals and conferences. He has served as a reviewer for many reputed journals. His research interests focus on the field of artificial intelligence with a special focus on NLP, machine learning, semantic Web, and learning technologies.

ALI SHARIQ IMRAN (Member, IEEE) received the master's degree in software engineering and computing from the National University of Science & Technology (NUST), Pakistan, in 2008, and the Ph.D. degree in computer science from the University of Oslo (UiO), Norway, in 2013. He is currently an Associate Professor with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway. He specializes in applied research with a focus on deep learning technology and its application to signal processing (both images and audio), natural language processing, and the semantic Web. He has over 65 peer-reviewed journals and conference publications to his name. He has served as a reviewer for many reputed journals over the years. He is a member of the Norwegian Colour and Visual Computing Laboratory (Colourlab) and the IEEE Norway Section. He is also an Associate Editor to open access journals, including IEEE ACCESS. He has served as a Guest Editor for Elsevier and Springer journals.

ARIANIT KURTI is currently an Associate Professor with the Department of Computer Science and Media Technology, Linnaeus University, Sweden. He is also a Senior Researcher with the RISE Research Institutes of Sweden. He has a strong applied research interest on interactive computing. His research focus has been on finding the most optimal approaches and methods for reaching sustainable digitalization by combining interaction design techniques, and AI and machine learning across different application domains both in society and industry. His recent research project focuses on applied machine learning approaches to personalize and improve learning trajectories of students in MOOC-like courses.

• • •