**IAES International Journal of Artificial Intelligence (IJ-AI)**
Vol. 15, No. 3, June 2026, pp. 2736~2749
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2736-2749 

# Hybrid deep learning for sentiment analysis of online student experiences

**Raja Ouadad, Hicham Mouncif**
Laboratory of LIMATI, Department of Mathematics and Informatics, Faculty of Polydisciplinary, Sultan Moulay Slimane University, Beni Mellal, Morocco

**Article Info**

***Article history:***

Received Feb 6, 2025
Revised Feb 12, 2026
Accepted Apr 22, 2026

***Keywords:***

COVID-19 pandemic
Hybrid deep learning model
Logistic regression
Online learning
Sentiment analysis

**ABSTRACT**

The COVID-19 pandemic disrupted millions of lives worldwide, and social media platforms became a significant outlet for people to share their emotions and experiences, providing valuable insights into the challenges and opportunities of remote education. This paper analyzes student sentiments about online learning during the pandemic using Twitter data. An experimental approach is developed to analyze public comments, focusing on the sentiment expressed in tweets related to online education. A hybrid deep learning model, based on the logistic regression (LR) sentiment model, is used to predict sentiment from a large dataset of online learning-related tweets. After performing n-gram analysis to extract key topics, tweets are classified into sentiment classes. The proposed convolutional long short-term memory (Conv-LSTM) and convolutional bidirectional long short-term memory (Conv-BiLSTM) models are trained on tweets annotated with granular sentiment classifications, achieving validation accuracies of 93% and 95%, respectively. This work provides meaningful insights into the emotional effects of online learning during the pandemic, contributing to the understanding of students' experiences and challenges in remote education.



***Corresponding Author:***

Raja Ouadad
Laboratory of LIMATI, Department of Mathematics and Informatics, Faculty of Polydisciplinary
Sultan Moulay Slimane University
Beni Mellal, Morocco
Email: ouadadraja2@gmail.com

## 1. INTRODUCTION

Originating in Wuhan, China, in December 2019, COVID-19 swiftly expanded beyond national borders, culminating in its recognition as a pandemic by the World Health Organization (WHO) on March 11, 2020 [1]. This unprecedented event severely impacted countries worldwide, resulting in millions of fatalities and necessitating strict lockdown measures to curb the virus's spread. Even with the advent of vaccines, maintaining social distancing remained crucial in mitigating transmission [2]. These restrictions brought significant changes to education, with traditional in-person learning transitioning to online platforms to ensure continuity. This shift underscored the critical role of e-learning as the primary alternative when physical attendance was not feasible [3]. Students faced considerable adjustments, adapting to an entirely new educational environment [4]. Their feedback during this transition has provided meaningful insights into the challenges and potential of online education, offering a basis for refining learning systems. Social media platforms, such as Twitter, have further enabled researchers to access and analyze student feedback more efficiently [5]. While opinions were traditionally gathered through questionnaires, the increasing prevalence of online platforms has facilitated the collection of diverse and spontaneous insights, enriched our understanding of students' experiences, and informed the continuous improvement of e-learning systems [6], [7].

Numerous studies have applied sentiment analysis techniques to analyze student feedback [8]. In these studies, feedback is often classified based on the sentiment polarity it conveys, categorizing responses as either positive or negative [9]. Both opinion mining and sentiment analysis share similar objectives, focusing on identifying sentiments, perspectives, and opinions embedded in texts produced by humans on a range of topics or entities [10]. The analysis relies on natural language processing (NLP), text mining, and feature extraction methods to infer and label sentiment orientation [11]. While closely related, sentiment analysis specifically targets emotional content, identifying words or phrases that reflect emotions, whereas opinion mining broadly extracts opinions about a particular subject or entity [7]. Over the years, researchers have explored various methods to accurately interpret sentiments in textual data. These approaches have progressed from lexicon-based techniques to machine learning algorithms, and more recently, to advanced hybrid deep learning models that leverage the combined strengths of multiple architectures to achieve improved performance in sentiment prediction [12].

In the context of our study, recent progress in hybrid deep learning models has demonstrated significant potential for analyzing student feedback and deriving meaningful insights into online learning experiences. These approaches combine the strengths of multiple neural network architectures to effectively manage the complexity and large volume of educational data. For example, Mary [13] introduced a hybrid model integrating convolutional neural networks (CNN) with long short-term memory (LSTM) networks, achieving high accuracy in evaluating student reviews from online learning platforms. This highlights the promise of deep learning techniques for processing unstructured feedback and supporting the improvement of teaching and learning practices. Similarly, Alawi and Bozkurt [14] integrated bidirectional encoder representations from transformers (BERT), bidirectional long short-term memory (BiLSTM), and CNN in a hybrid model to analyze Turkish tweets about university satisfaction, achieving an accuracy of over 91% and effectively handling the linguistic complexities of Turkish. Expanding this research into massive open online courses (MOOCs), Mrhar *et al.* [15] utilized a Bayesian CNN-LSTM model to analyze 140,320 Coursera course reviews, achieving 91% accuracy while linking forum post sentiments to dropout risks and course success. Complementing this, Li *et al.* [16] presented a shallow hybrid BERT–CNN approach for sentiment analysis based on 19,148 comments extracted from a Chinese MOOC platform, achieving 81.3% accuracy and 92.8% F1-score, all while significantly reducing computational complexity. Similarly, Zheng *et al.* [17] introduced a model combining BERT and a bidirectional gated recurrent unit (BiGRU) model within the feedback system structure (FSS) for intelligent teacher-student interaction and curriculum enhancement. On a self-constructed dataset of online education platform reviews, their model achieved a remarkable accuracy of 98.82%, surpassing traditional methods like naive Bayes by 21.54%. Additionally, their model demonstrated strong sentiment analysis capabilities, especially in handling texts containing Chinese internet buzzwords, highlighting its effectiveness in intelligent feedback systems for educational contexts. In alignment with these advancements, Jebbari *et al.* [18] presented a hybrid approach for sentiment analysis in MOOC forums that merges word-level representations from BiLSTM layers with character-level features from CNN layers to effectively encode unique linguistic patterns. With an accuracy of 93.11%, the model highlights its effectiveness and contributes significantly to sentiment analysis in online education platforms. Collectively, these studies illustrate the transformative potential of hybrid deep learning approaches in educational sentiment analysis across diverse contexts and datasets.

Building on the insights gained from analyzing public sentiment related to the shift toward online education prompted by the COVID-19 pandemic, this study introduces several key contributions to advance sentiment analysis techniques. These contributions include leveraging lexical n-gram models to identify prominent trends in tweet-specific vocabulary and employing domain-specific word analysis to measure the impact of online learning-related terms. The sentiment polarity ratings were fine-tuned using a logistic regression (LR) model, facilitating the classification of tweets into positive, negative, and neutral categories. Furthermore, a hybrid convolutional long short-term memory (Conv-LSTM) model was developed, leveraging these refined ratings to predict sentiment trends more accurately. Together, these methods provide a robust framework for understanding and addressing the evolving challenges of online learning in times of global disruption.

The paper is structured as follows: section 2 introduces the study's methods, commencing with a description of the proposed methodology and the organization of the dataset, preprocessing techniques, and the identification of word trends using n-gram models. It also covers the detection of online learning-specific terms during the COVID-19 pandemic, sentiment analysis and classification using natural language toolkit (NLTK), the development process of the sentiment classification model using LR, analysis of global sentiment trends, and the architecture and hyperparameters of the hybrid Conv-LSTM model. Section 3 discusses the model's performance and compares it with other state of art models. The final section summarizes the findings and explores potential directions for future research.

## 2. METHOD

### 2.1. Overview

The COVID-19 pandemic has profoundly transformed education on a global scale, shifting conventional in-person learning to online learning platforms because of lockdowns and social distancing measures. Although this sudden shift posed challenges, such as adapting to new technologies and managing self-directed learning, it also highlighted the potential of online education as a more flexible and accessible alternative. To assess its effectiveness during this period, our research examined learner feedback by analyzing a dataset of 310,547 tweets related to online learning and courses, posted between January and October 2020. These tweets underwent a thorough cleaning and preprocessing process using the NLTK package. This included removing noisy data, tokenizing text into linguistic units, and applying lemmatization to reduce words to their fundamental lexical forms for more in-depth analysis. After the cleaning and preprocessing step, we identified the most frequently occurring unigrams, bigrams, and trigrams across the entire tweet dataset. Subsequently, we identified terms specific to the effects of the COVID-19 pandemic on online education, as well as their frequency in the reviews, to better understand their influence on learners' opinions.

Additionally, for each cleaned and preprocessed tweet, the polarity was assessed using a sentiment analysis tool from the NLTK library. Utilizing an adaptive classification approach, tweets were categorized into three sentiment classes: positive, negative, and neutral. An LR model was then developed, leveraging the most frequent n-grams to achieve fine-grained sentiment classification of tweets. This model assigns a detailed sentiment score to each review by computing probability scores, ensuring accurate classification into the appropriate sentiment categories. We then assigned sentiment scores to reviews in five distinct classes, ranging from the highly positive (1.0) to the highly negative (-1.0), based on the obtained probability scores. Neutral reviews were grouped into a separate class (0.0). Finally, we examined the daily variations in overall sentiment trends among learners about online courses during the COVID-19 period by calculating the average sentiment for each day. The tweets, labeled on a fine-grained scale from 0 to 4, were transformed into word embeddings to facilitate further analysis. The dataset was partitioned, allocating 80% of the samples for training and 20% for validation of the hybrid models. Both sequential and BiLSTM neural networks were trained on the vectorized tweets and their corresponding labels to predict sentiment polarity. Following the training process, the models' performance was evaluated on test datasets to assess their accuracy. The overall system architecture adopted in this study is summarized in Figure 1.

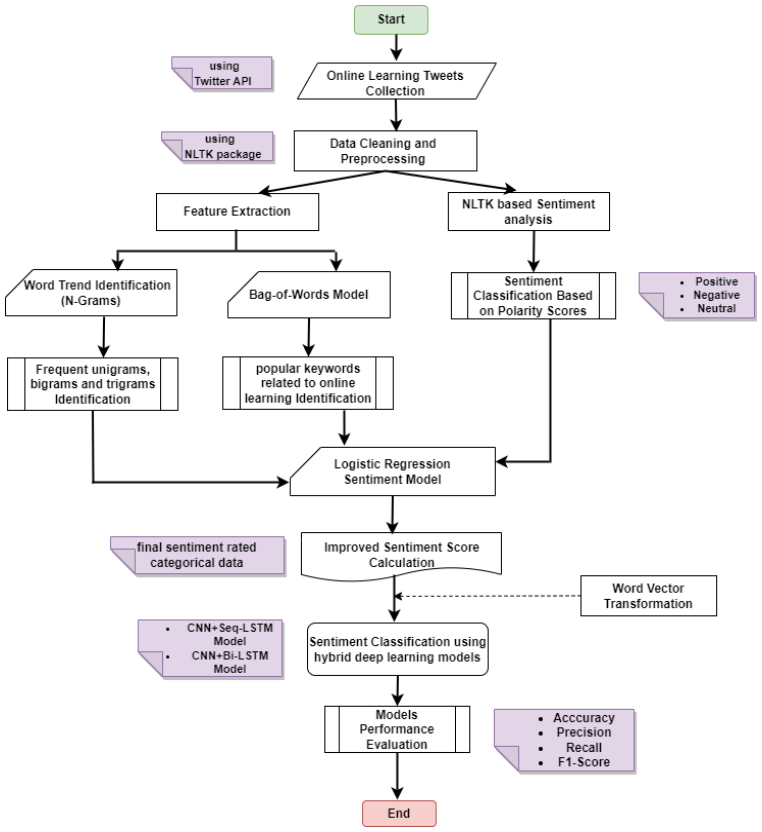



Figure 1. Proposed system architecture

### 2.2. Dataset description

The dataset employed in this study was collected through the Twitter API, which necessitates the use of a Twitter developer account. A custom Python script, get_tweets.py, was developed to collect English-language tweets related to distance learning. The script searched for tweets containing the following hashtags: #distancelearning, #onlineschool, #onlineteaching, #virtuallearning, #onlineducation, #distanceeducation, #OnlineClasses, #DigitalLearning, #elearning, and #onlinelearning. Additionally, the script targeted keywords such as "distance learning," "online teaching," "online education," "online course," "online semester," "distance course," "distance education," "online class," and "e-learning." To maintain data quality and prevent redundancy, retweets were removed from the dataset.

The get_tweets function temporarily stored the retrieved tweets in a panda DataFrame and saved the data as CSV files in an output directory. The data collection process took approximately 45 hours to gather 326,140 tweets. Subsequently, the concatenate.py script was employed to merge all individual CSV files into a unified dataset. The final dataset, referred to as tweets_raw.csv, contains 310,547 tweets, with various features, as outlined in Table 1.

### 2.3. Data preprocessing

Empirical data is inherently imperfect, often containing inconsistencies or irrelevant information. The preprocessing stage is vital for addressing these issues by removing noise and irrelevant elements, thereby improving data quality and ensuring optimal performance of machine learning models [16]. In NLP, text preprocessing plays a vital role in organizing and refining textual data, which in turn improves data quality and the effectiveness of later analyses. The current study focuses on tweets related to online learning during the COVID-19 pandemic, which were collected using the Twitter API. The raw data, which contained redundant information and duplicate entries, was processed and cleaned to ensure its suitability for analysis. During the data collection phase, we initially gathered approximately 326,140 tweets. To ensure data quality, all duplicate tweets were removed from the dataset, resulting in a final dataset of 310,547 unique tweets. The dataset was then preprocessed using a user-defined function built with NLTK, a widely used Python toolkit for NLP.

The preprocessing procedure involved standardizing the textual data by converting all tweets to lowercase. Additional cleaning steps included removing extra spaces, numeric values, special symbols, URLs, punctuation, and commonly used stop words. Mentions and hashtags were also eliminated to avoid introducing bias during feature extraction. The tweets were then tokenized, splitting the text into individual words or meaningful phrases. Lemmatization was applied to convert words into their base forms (lemmas), ensuring that different forms of a word (e.g., "studying," "studied," and "studies") were treated as a single representation, which helped normalize the dataset. These preprocessing steps effectively prepared the tweets for subsequent analysis. Figure 2 illustrates the daily distribution of tweets, highlighting submission frequencies over time. This visualization provides insights into temporal patterns, revealing periods of heightened or reduced activity in the dataset.

Table 1. Features and data types of the collected tweet dataset

| Feature | Description | Data type |
|---|---|---|
| Id | Unique identifier for each tweet | Number |
| Content | Text of the tweet | Text |
| Location | Location of the user (if available) | Text |
| Username | Twitter username of the tweet author | Text |
| Retweet count | Number of retweets | Number |
| Favorite | Number of likes or favorites | Number |
| Created at | Timestamp when the tweet was posted | DateTime |

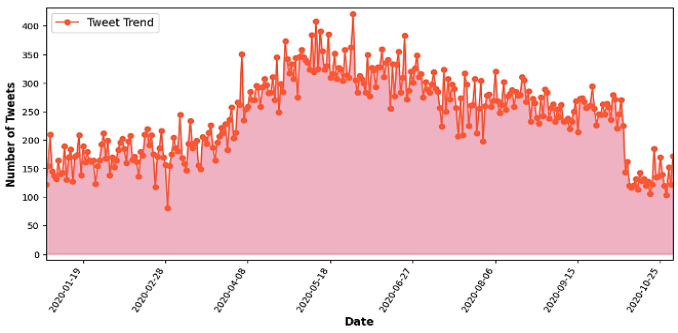



Figure 2. Temporal distribution of COVID-19 tweets in the corpus

### 2.4. Identifying word trends with n-grams

In the field of NLP, n-gram models are frequently applied to analyze statistical patterns and capture syntactic structures [19]. To explore the tokenized corpus, we computed the frequency of single words and adjacent word combinations through an n-gram approach. We then applied the chain rule of probability to estimate the likelihood of specific word sequences appearing, as defined in (1).

$$P(x^1, x^2, x^3, \dots, x_n) = P(x^1)P(x^2|x^1)P(x^3|x^1, x^2) \dots P(x_n|x^1, x^2, x^3, \dots, x_n{}^{-1}) = \prod_{i=1}^{n} P(x_i \mid x_1^{i-1}) \quad (1)$$

For example, a sentence can be considered as "online learning is expanding". Based on the chain rule of probability, the probability of the sentence can be expressed as: P ("online learning is expanding") = P("online") × P ("learning" | "online") × P ("is" | "online learning") × P ("expanding" | "online learning is"). In (2), the chain rule of probability is used to determine the probabilities of words in each sentence.

$$P(W_1^n) = \prod_l P(W_j \mid W_1, W_2, W_3, \dots, W_{j-1}) = \prod_{j=1}^{n} P(W_j \mid W_1^{j-1}) \quad (2)$$

The principle of the Markov assumption asserts that a word's occurrence is conditioned solely on the word directly before it. This model allows the prediction of a future word based on the immediate past, rather than relying on the entire sequence of previous words. In the bigram model, the probability of a word is computed using the conditional probability $P(Wi|Wi-1)$, which takes into account only the preceding word, without considering any other prior words [20] as in (3).

$$P(W_1, W_2) = \prod_{(i=2)} P(W_2 \mid W_1) \quad (3)$$

The formula for calculating the bigram probability is (4).

$$P(W_k|W_{k-1}) = \frac{count(W_{(k-1)}, W_k)}{count(W_{(k-1)})} \quad (4)$$

The n-gram model was used to extract the most common unigrams, bigrams, and trigrams from the dataset. This analysis sheds light on prominent topics, uncovering the key drivers of sentiment and providing deeper insights into human emotions expressed in tweets about online courses and learning during the COVID-19 pandemic. Among these, unigrams appeared most frequently, followed by bigrams, with trigrams being the least common. Figure 3 shows the top 30 unigrams, bigrams, and trigrams, along with their respective occurrence counts.

### 2.5. Online learning-specific word detection during COVID-19

Following the completion of the preprocessing stage, a bag-of-words (BoW) model was developed using the most commonly occurring terms from the dataset, highlighting keywords related to online learning and education during the COVID-19 pandemic. Additionally, a word cloud was generated to provide a visually appealing representation of the extracted text data, as shown in Figure 4. Several words have been recognized from the curated lexicon, occurring at different points and in diverse contexts throughout the tweets. Analyzing the highest frequency occurring words in the data set provides valuable insights into their influence on the overall sentiment of the tweets. Approximately 9,871 million words were identified from the dataset. To pinpoint the most prominent words concerning online courses and learning during the COVID-19 pandemic, the frequency of each token was assessed. A popularity score for each word was then computed, offering a quantitative measure of its prominence within the dataset.

After identifying the word frequencies, the likelihood of occurrence for each word was calculated based on the total of words in the dataset. Table 2 shows the prominence and associated probability measures of the words appearing most frequently in the dataset. The probability of each word in the dataset can be calculated using (5), which normalizes the frequency of each word by dividing its count by the total word count in the dataset.

$$P(W_i) = \frac{count(W_i)}{\sum_{i=0}^{n} count(W_{i=0}^{n})} \quad (5)$$

This normalization method provides a clearer perspective on the relative frequency of words in the dataset, helping to identify key themes. For instance, the word "online" stands out with the highest

frequency—134,112 occurrences—representing a probability of 0.027 (2.71%). Similarly, “course” and “learning” follow with 112,226 occurrences (2.26%) and 64,367 occurrences (1.30%), respectively. These prominent figures underline the strong emphasis on online platforms and educational content within the dataset. Other terms, such as “class” (61,103 occurrences, 1.23%) and “student” (41,475 occurrences, 0.84%), further reinforce the learner-centric nature of the discussions. Collectively, these insights reveal the central focus on online learning and course experiences in the analyzed content.

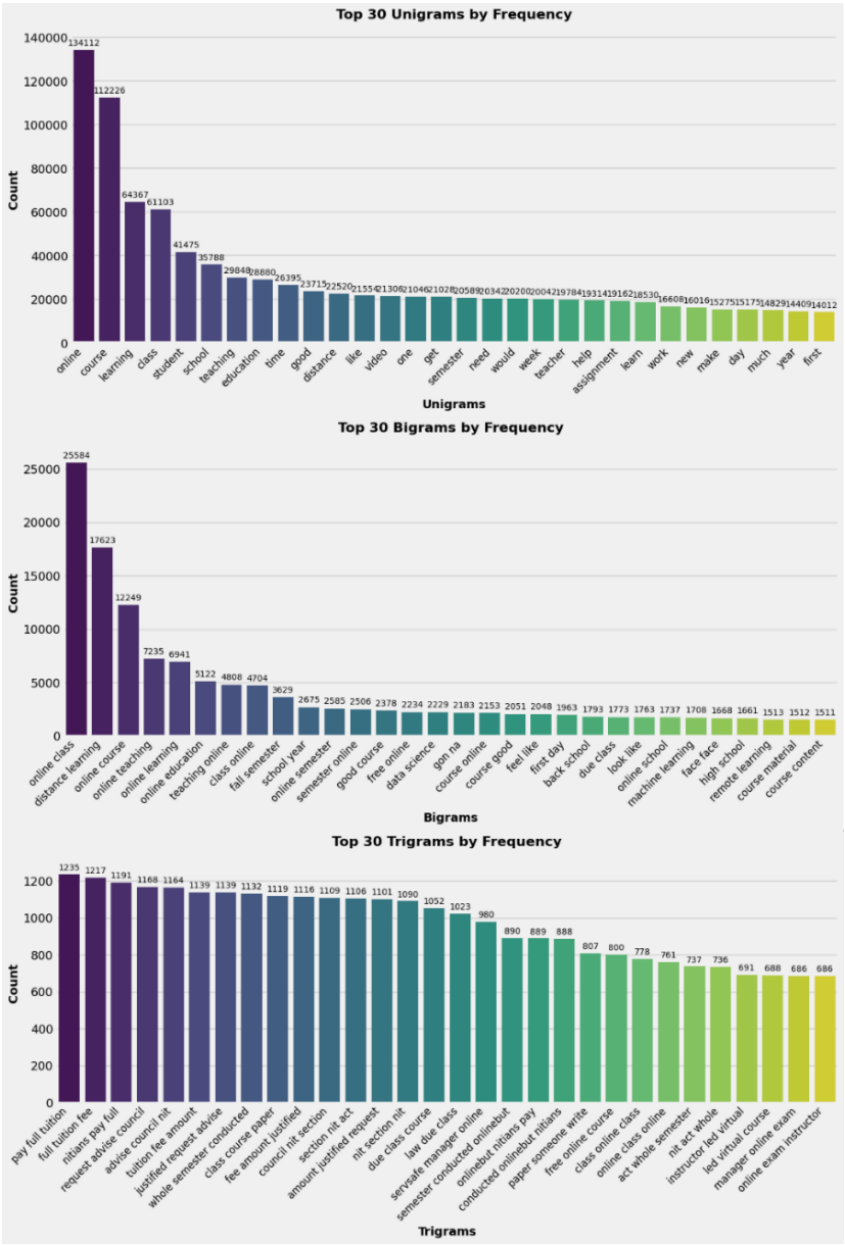



Figure 3. Top 30 unigrams, bigrams, and trigrams with their occurrence counts

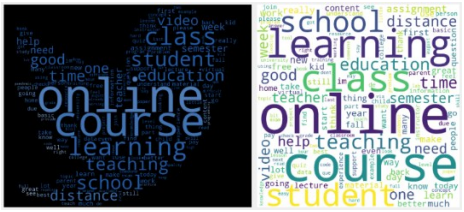



Figure 4. Word clouds highlighting key terms in online learning during the pandemic

Table 2. Frequency and probability of highly occurring words

| Word | Popularity | Probability |
|---|---|---|
| Online | 134,112 | 0.027060 |
| Course | 112,226 | 0.022644 |
| Learning | 64,367 | 0.012987 |
| Class | 61,103 | 0.012329 |
| Student | 41,475 | 0.008369 |

### 2.6. Sentiment analysis and classification using natural language toolkit

In recent years, deep learning-based sentiment analysis has gained significant popularity for deriving insights and predicting text to evaluate public responses to different events [21]. This method is mainly employed to monitor shifts in public opinion and sentiment. It belongs to the field of data mining and utilizes computational linguistics and strategies for the examination and analysis of text. By examining user-generated subjective information on social networking platforms, sentiment analysis allows for the classification of text into different sentiment categories, offering valuable insights into public feelings and attitudes [22].

The NLTK sentiment analyzer is a widely used tool for evaluating sentiment polarity. Renowned for providing reliable and accurate results in text analysis, the sentiment analysis functionality from the NLTK library was employed in this study to evaluate the sentiment polarities of preprocessed tweets. This tool applies predefined rules and heuristics to perform a detailed sentiment assessment. Given that the dataset consists of raw tweets without labeled sentiment classes, sentiment analysis is essential for understanding the distribution of tweets across various sentiment categories. Each tweet's sentiment polarity is computed using NLTK's sentiment analyzer, which generates scores for positive, negative, and neutral sentiments, as well as an overall compound sentiment score.

The overall sentiment (compound score) is obtained by summing the valence scores of the words present in the lexicon, which are then adjusted according to a set of predefined rules. These scores are normalized to a range between -1 (extremely negative) and +1 (extremely positive), providing a single, unidimensional measure of sentiment for the entire sentence. A compound score closer to +1 indicates a more positive sentiment, while a score closer to -1 reflects a more negative sentiment. Tweets are classified into three sentiment categories—positive, negative, or neutral—based on their compound polarity scores. If the polarity score is greater than 0, the tweet is considered positive; if less than 0, it's negative; and if the score equals 0, the tweet is deemed neutral. The sentiment classification of our dataset is visually represented in Figure 5, where tweets are categorized into positive, neutral, and negative sentiments. The figure shows that 59.31% of the tweets express a positive sentiment, while 23.08% are neutral, and 17.61% reflect negative sentiment. Positive and negative tweets are further analyzed for a more granular sentiment assessment, providing a deeper understanding of the overall sentiment distribution in relation to online learning during the COVID-19 pandemic.

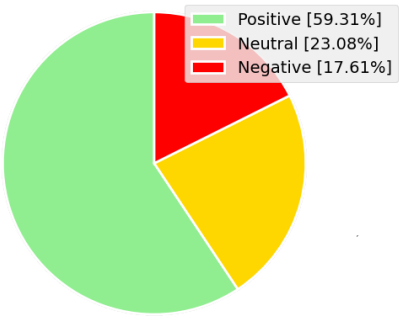



Figure 5. Sentiment distribution of tweets using NLTK

### 2.7. Development of the logistic regression sentiment model

LR is a widely used statistical model for binary classification tasks, making it a suitable choice for sentiment analysis [23]. It is particularly effective in high-dimensional spaces, such as text data, and offers interpretability through its probabilistic predictions. Unlike linear regression, which produces continuous outputs, LR applies the sigmoid function to map the weighted sum of input features to a probability score between 0 and 1 [24]. The probability of an instance belonging to the positive sentiment class (Y =1) is computed as in (6).

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(wX+b)}} \quad (6)$$

Where $X$ denotes the feature vector derived from the input text, $w$ represents the weight coefficients learned during the training process, $b$ is the bias term, and $e$ corresponds to Euler's number.

In this study, we employ a sentiment classification model based on LR to refine the polarity ratings of tweets. LR, a discriminative model, is chosen for its computational efficiency, interpretability, and ability to handle large-scale textual data [25]. The model is optimized using maximum likelihood estimation (MLE), ensuring an effective decision boundary between sentiment classes. During feature extraction, we apply n-gram modeling to capture unigrams, bigrams, and trigrams, alongside domain-specific words from the BoW model. The extracted features are transformed using term frequency-inverse document frequency (TF-IDF) to normalize term importance, reducing the influence of high-frequency words. The model is trained on a dataset comprising 320,000 textual tweets, using the lbfgs solver with 1,000 iterations to ensure convergence in high-dimensional spaces. Each new tweet is transformed into a TF-IDF feature vector and passed through the logistic function to compute the predicted sentiment class and its associated probability score, providing probabilistic confidence levels essential for sentiment classification applications.

The classifier predicts the probability of each tweet belonging to the positive or negative sentiment class, with scores below 0.5 reflecting negative sentiment and scores above 0.5 reflecting positive sentiment. Using the classification algorithm presented in Algorithm 1, the tweets are further categorized into highly positive (≥0.75), positive (<0.75), negative (>0.25), and highly negative (≤0.25). The sentiment classification results are further refined by incorporating neutral tweets with sentiment scores derived from an alternative classification approach. As illustrated in Figure 6, the refined sentiment distribution shows 52.57% of tweets classified as highly positive, 9.93% as positive, 22.80% as neutral, 10.18% as negative, and 4.52% as highly negative. Given its scalability, interpretability, and probabilistic output, LR is well-suited for large-scale sentiment analysis tasks, offering robust and efficient sentiment prediction across diverse textual datasets.

Algorithm 1. Fine-grained sentiment classification

```
Input: dataset of tweets with:
-    class (positive, neutral, or negative).
-    LR_proba (probability score).
Output: refined_sentiment (fine-grained sentiment):
-    1.0: highly positive
-    0.5: positive
-    0.0: neutral
-    -0.5: negative
-    -1.0: highly negative
For each tweet i in the dataset:
     If the tweet class is positive and score (LR_proba) >=0.75:
          Assign refined_sentiment =1.0
     Else if the tweet's class is positive and score (LR_proba) <0.75:
          Assign refined_sentiment =0.5
     Else if the tweet's class is neutral:
          Assign refined_sentiment =0.0
     Else if the tweet's class is negative and score (LR_proba)>0.25:
          Assign refined_sentiment =-0.5
     Else:
          Assign refined_sentiment =-1.0
End loop.
```

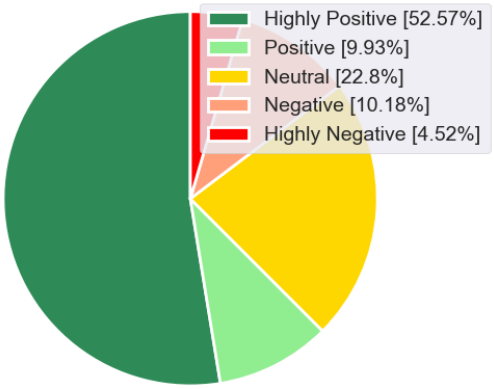



Figure 6. Sentiment distribution of tweets using the LR sentiment model

### 2.8. Analysis of global sentiment trends

In this analysis, sentiment-rated evaluations with a high degree of precision, along with their refined sentiment scores, were used to examine sentiment trends in online courses. Representative reviews for each class of sentiment, as well as their associated LR probability scores, are shown in Table 3. These reviews capture a range of emotions, from highly positive to highly negative, which contribute to fluctuations in the global sentiment trend. By calculating date-wise average sentiment scores, notable variations over time were identified. These variations provide valuable insights into the dynamic emotional landscape of learners. The positive, neutral, and negative sentiments revealed in the reviews highlight key moments of emotional change, offering a deeper understanding of how learners' experiences evolve in response to the online learning environment.

Table 3. Online courses tweets from different sentiment classes

| Date | Original Tweet | Class | Prob | Refined sentiment | Refined rating |
|---|---|---|---|---|---|
| 2020-05-22 | 90% of graduates have achieved their goals. Join them! | pos | 0.9476528345971169 | Highly positive | 1.0 |
| 2020-04-26 | Distance learning forced parents to participate in their kids' education. I think it was actually a good thing. | pos | 0.7472851765195649 | Positive | 0.5 |
| 2020-08-03 | So, I can't believe our online class was recorded that means my parents will see how i act in class. | neu | 1.000000 | Neutral | 0.0 |
| 2020-09-01 | Very low-level content. No strategies or actionable lessons. | neg | 0.47437446874184913 | Negative | -0.5 |
| 2020-10-08 | Teachers now have to compete with social media as a source of information. The distance learning setup makes it difficult for them to guide students immersed in false information online. | neg | 0.053669205380508685 | Highly negative | -1.0 |

### 2.9. Architecture and hyperparameters of the hybrid convolutional long short-term memory model

The LSTM network has proven to be highly effective in sentiment analysis, surpassing other neural network models in terms of prediction accuracy. LSTM, a specialized type of recurrent neural network (RNN), is capable of forecasting future outcomes based on features extracted from the dataset [26]. Its architecture consists of memory cells ($c_t$) and three gates: input ($i_t$), forget ($f_t$), and output ($o_t$). These gates control the flow of information by allowing new data into the memory cell, forgetting outdated information, and propagating the current memory state through the network. The LSTM effectively prevents the vanishing gradient problem, ensuring accurate predictions. The gates and memory cells are governed by weight matrices ($W$), the logistic sigmoid activation function (σ), and the Hadamard product ($\odot$), which together regulate the flow and retention of information [27]. The mathematical formulation governing the behavior of the LSTM gates, memory cell state, and hidden state is defined in (7).

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}x_{t-1} + W_{ci}K \odot c_{t-1} + b_i) \\
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}K \odot c_{t-1} + b_f\right) \\
c_t &= f_t \odot Kc_{t-1} + i_t \odot Ktanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}Kc_t + b_o) \\
h_t &= o_t \odot Ktanh(c_t)
\end{aligned} \tag{7}
$$

The proposed hybrid Conv-LSTM model effectively addresses the redundancy issues often encountered by LSTM networks. While LSTMs are proficient at capturing local dependencies between neighboring words in large datasets, they face challenges in optimizing performance. The Conv-LSTM model combines convolutional layers with LSTMs to overcome these issues, providing improved accuracy in sentiment classification.

This study tweets were labeled into sentiment classes ranging from most negative (0) to most positive (4), with the dataset divided into 80% for training and 20% for testing. Each tweet was preprocessed and transformed into 100-dimensional word embeddings using the embedding layer, which was configured with vocabulary size of 87,587 unique words and maximum sequence length of 460. Sentiment classification was performed using both sequential and BiLSTM models, implemented with TensorFlow and Keras. The model architecture incorporated embedding, convolutional, max pooling, LSTM, and dense layers.

The network architecture, illustrated in Figure 7, starts with an embedding layer that transforms the input text into dense vectors of 100 dimensions. This is followed by a one-dimensional convolutional layer containing 128 filters with a kernel size of 3, using the ReLU activation function. To stabilize the training process, batch normalization was applied, and a MaxPooling layer with a pool size of 2 was used to downsample the feature maps. A dropout rate of 0.3 was incorporated within the convolutional layers to prevent overfitting.

The outputs generated by the convolutional layers are fed into a LSTM layer comprising 128 units, which enables the model to capture long-range dependencies within the input sequences. To prevent overfitting, a dropout rate of 0.4 was applied to the LSTM layer. Instead of flattening the sequence outputs, a GlobalMaxPooling1D layer was used to identify and retain the most salient features. The resulting feature representation is then passed through two fully connected dense layers with 64 and 32 units, respectively, each followed by dropout rates of 0.3 and 0.2 to further enhance regularization.

The final layer consists of five units with softmax activation, enabling multiclass sentiment classification. Compilation of the model involved the Adam optimizer with a 0.001 learning rate, while the loss was computed using categorical cross-entropy. To dynamically adjust the learning rate when validation loss plateaus, the ReduceLROnPlateau callback was used, reducing the rate by a factor of 0.5 with a patience of three epochs. The complete model, which contains 10,815,281 trainable parameters, demonstrates a strong capacity to learn from fine-grained textual data, resulting in enhanced sentiment prediction performance, particularly for tweets related to online learning.

Model: "sequential_1"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| embedding_1 (Embedding) | (None, 460, 100) | 8,758,700 |
| conv1d_1 (Conv1D) | (None, 460, 128) | 38,528 |
| max_pooling1d_1 (MaxPooling1D) | (None, 230, 128) | 0 |
| dropout_3 (Dropout) | (None, 230, 128) | 0 |
| lstm_1 (LSTM) | (None, 230, 128) | 131,584 |
| dropout_4 (Dropout) | (None, 230, 128) | 0 |
| flatten_1 (Flatten) | (None, 29440) | 0 |
| dense_3 (Dense) | (None, 64) | 1,884,224 |
| dropout_5 (Dropout) | (None, 64) | 0 |
| dense_4 (Dense) | (None, 32) | 2,080 |
| dense_5 (Dense) | (None, 5) | 165 |

Total params: 10,815,281 (41.26 MB)
Trainable params: 10,815,281 (41.26 MB)
Non-trainable params: 0 (0.00 B)

Figure 7. Architecture of the proposed hybrid Conv-LSTM model

## 3. RESULTS AND DISCUSSION

This section reports the performance outcomes of the proposed hybrid model for classifying the sentiment of tweets associated with online learning and courses. Model evaluation was conducted using standard metrics, including accuracy, precision, F1-score, and support. To assess its effectiveness, the results of the proposed approach are compared with those of other state-of-the-art models previously applied to student reviews in online learning contexts, highlighting the improvements achieved in sentiment classification accuracy.

### 3.1. Model performance

During training, both the hybrid Conv-LSTM and convolutional bidirectional long short-term memory (Conv-BiLSTM) models were trained with identical conditions: a batch size of 32, a learning rate of 0.001, and a verbosity level of 2. To enhance generalization and reduce overfitting, the dropout technique was applied. Table 4 reports the training and validation accuracy and loss values of the proposed hybrid models across six training epochs. The Conv-LSTM model achieved a peak training accuracy of 95.78% with

a corresponding training loss of 0.1617, while its validation accuracy reached 93.32% with a validation loss of 0.2259. Similarly, the Conv-BiLSTM model demonstrated superior performance, achieving a final training accuracy of 98.55% with a training loss of 0.0165 and a validation accuracy of 95.20% with a validation loss of 0.0639. The enhanced performance of the Conv-BiLSTM model can be attributed to its bidirectional architecture, enabling it to capture information from both past and future elements of the sequence to gain a deeper insight into the input data.

The classification reports in Table 5 further confirm the Conv-BiLSTM model's better performance. It consistently delivered higher precision, recall, and F1-scores across all categories, especially excelling in the "negative" and "positive" sentiment classes. While the Conv-LSTM model recorded strong macro and weighted averages (0.88 and 0.93, respectively), the Conv-BiLSTM achieved noticeably higher scores (macro average: 0.91 and weighted average: 0.95), showcasing its strong ability to handle diverse sentiment categories effectively.

One key observation is the slight overfitting observed in the Conv-LSTM model, evident from the gap between its training and validation losses. This might be due to periodic sentiment variations in the dataset, possibly caused by changes in context and timing during data collection. Despite this, both models successfully captured sentiment patterns, with the Conv-BiLSTM emerging as the more reliable and accurate architecture for sentiment analysis, and confirmed that incorporating bi-directional processing enhances sentiment classification, particularly for nuanced emotional expressions.

Table 4. Training and validation accuracy and loss for hybrid models

| Models | Epochs | Training accuracy | Training loss | Validation accuracy | Validation loss |
|---|---|---|---|---|---|
| Conv-LSTM model | 1/6 | 0.8233 | 0.5904 | 0.9022 | 0.3027 |
| | 2/6 | 0.9097 | 0.2941 | 0.9187 | 0.2556 |
| | 3/6 | 0.9322 | 0.2346 | 0.9268 | 0.2422 |
| | 4/6 | 0.9434 | 0.2014 | 0.9268 | 0.2417 |
| | 5/6 | 0.9528 | 0.1761 | 0.9330 | 0.2302 |
| | 6/6 | 0.9578 | 0.1617 | 0.9332 | 0.2259 |
| Conv-BiLSTM model | 1/6 | 0.8903 | 0.1092 | 0.9309 | 0.0690 |
| | 2/6 | 0.9468 | 0.0538 | 0.9406 | 0.0624 |
| | 3/6 | 0.9657 | 0.0366 | 0.9451 | 0.0595 |
| | 4/6 | 0.9764 | 0.0264 | 0.9487 | 0.0570 |
| | 5/6 | 0.9823 | 0.0203 | 0.9511 | 0.0611 |
| | 6/6 | 0.9855 | 0.0165 | 0.9520 | 0.0639 |

Table 5. Classification report for hybrid models

| Class | Conv-LSTM model | | | | Conv-BiLSTM Model | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision (%) | Recall (%) | F1-score (%) | Support | Precision (%) | Recall (%) | F1-score (%) | Support |
| Highly negative | 86 | 93 | 89 | 2,896 | 93 | 97 | 95 | 965 |
| Negative | 85 | 83 | 84 | 6,402 | 87 | 80 | 83 | 2,411 |
| Neutral | 99 | 97 | 98 | 13,962 | 99 | 97 | 98 | 13,962 |
| Positive | 73 | 73 | 73 | 6,054 | 78 | 81 | 79 | 6,117 |
| Highly positive | 97 | 97 | 9 | 32,568 | 97 | 98 | 97 | 38,427 |
| Accuracy | - | - | 93 | 61,882 | - | - | 95 | 61,882 |
| Macro avg | 88 | 89 | 88 | 61,882 | 91 | 90 | 91 | 61,882 |
| Weighted avg | 93 | 93 | 93 | 61,882 | 95 | 95 | 95 | 61,882 |

The evaluation using confusion matrices (Figure 8) shows that while both models perform well, the Conv-BiLSTM model outperforms the Conv-LSTM, particularly when handling less common sentiment classes like "negative" and "positive". This demonstrates the Bi-LSTM's ability to capture bidirectional dependencies and extract richer contextual patterns from the data. The Conv-BiLSTM model proves to be more robust and effective, making it the superior hybrid architecture for sentiment analysis. The model's capacity to manage a wide range of sentiment patterns effectively implies its potential to support educators, course designers, and administrators in making data-driven decisions to enhance the quality of online education. Additionally, the observed overfitting in the Conv-LSTM model suggests that sentiment in educational feedback may vary due to contextual and temporal factors, emphasizing the need for models that generalize well. By leveraging the bidirectional architecture's strength in contextual understanding, institutions can better analyze student sentiment, adapt course content, and personalize learning experiences, ultimately contributing to improved learner satisfaction and engagement in online education. These findings highlight the importance of integrating advanced sentiment analysis tools into educational platforms to drive meaningful and targeted improvements.

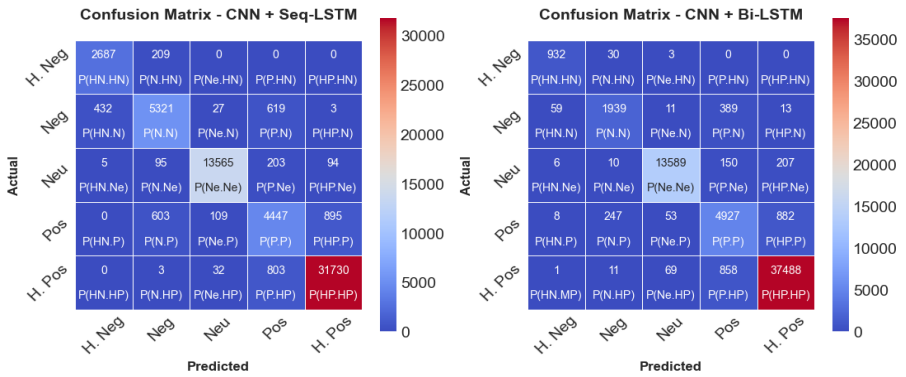



Figure 8. Confusion matrix for Conv-LSTM and Conv-BiLSTM models

### 3.2. Performance comparison with other state-of-the-art models

This section evaluates the performance of the proposed Conv-BiLSTM model in comparison with existing hybrid deep learning approaches applied to sentiment analysis tasks on online learning–related datasets. Table 6 presents a comparative performance analysis between the proposed model and several state-of-the-art deep neural network architectures reported in the literature. The proposed Conv-BiLSTM model achieves an accuracy of 95% on a large-scale dataset comprising 310,547 tweets related to online learning, outperforming several hybrid models such as BERT+CNN [16], Hybrid CNN-LSTM [28], and Bayesian CNN+LSTM [15].

The observed performance gains are consistent with previous studies that highlight the benefits of integrating convolutional layers with bidirectional recurrent layers for sentiment analysis. For instance, models such as CNN-BLSTM-AT [29] and BERT+BiLSTM+CNN [14] also report improved classification accuracy by leveraging both local feature extraction and long-term contextual modeling. The superior performance of the proposed Conv-BiLSTM model can be attributed to its bidirectional architecture, which captures context provided by tokens occurring before and after a given position, as well as the convolutional layers' ability to extract discriminative n-gram features. Although some studies report competitive performance on smaller or domain-specific datasets, the results obtained in this work demonstrate strong generalization capability on a significantly larger corpus. This indicates that the proposed model is particularly effective in handling large-scale, real-world textual data related to online learning and course evaluations, thereby validating its robustness and practical applicability.

Table 6. Comparison of sentiment classification accuracy across deep neural network models

| Deep neural network classifier | Dataset description | Sentiment classification accuracy (%) |
|---|---|---|
| Our model | 310,547 tweets related to online learning | 95 |
| BERT+CNN [16] | 19,148 comments from a Chinese MOOC platform | 81.3 |
| Hybrid CNN-LSTM [28] | 59,391 Google Play reviews from 9 educational platforms (Duolingo, Coursera, and Udemy) | 93.28 |
| Final hidden states of LSTM+Dependency tree-LSTM+SVM [30] | Vietnamese students' feedback corpus (16,000 sentences) | 90.7 |
| BERT+BiLSTM+CNN [14] | 7,793 tweets about university satisfaction | 91 |
| Bayesian CNN+LSTM [15] | 140,320 Coursera course reviews | 91 |
| CNN-BLSTM-AT [29] | 10,000 comments from online teaching evaluations | 93 |
| BiLSTM(word-level)+CNN (character-level) [18] | 24,684 anonymized posts from the Stanford MOOC forum | 93.11 |

## 4. CONCLUSION

In summary, this study highlights the significance of examining students' sentiments regarding online learning throughout the COVID-19 pandemic. By employing convo-sequential and Conv-BiLSTM networks, we classified tweets related to online education into distinct sentiment categories, achieving high validation accuracy and providing valuable insights into the challenges students faced during the shift to remote learning. This work sets the stage for further exploration of sentiment analysis in education, with the potential to improve online learning environments. Future research could expand the dataset to include diverse social media platforms, examine factors like course type and instructor quality, and refine sentiment models using transformer-based approaches such as BERT, ultimately enabling more personalized and adaptive online learning experiences based on emotional insights.

**FUNDING INFORMATION**

Authors state no funding involved.

**AUTHOR CONTRIBUTIONS STATEMENT**

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Raja Ouadad | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | |
| Hicham Mouncif | | | | | ✓ | | ✓ | | | ✓ | | ✓ | | ✓ |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

**CONFLICT OF INTEREST STATEMENT**

Authors state no conflict of interest.

**INFORMED CONSENT**

We have obtained informed consent from all individuals included in this study.

**ETHICAL APPROVAL**

The research related to human use has been compiled with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

**DATA AVAILABILITY**

The data that support the findings of this study are available from the corresponding author, [RO], upon reasonable request.

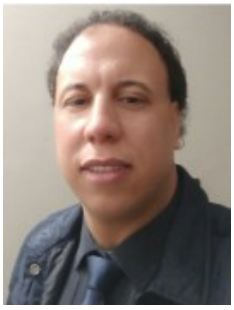
**REFERENCES**

[1] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing," *Internet of Things*, vol. 11, 2020, doi: 10.1016/j.iot.2020.100222.
[2] A. D. Dubey, "Twitter sentiment analysis during COVID19 outbreak," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3572023.
[3] D. J. Lemay, P. Bazelais, and T. Doleck, "Transition to online learning during the COVID-19 pandemic," *Computers in Human Behavior Reports*, vol. 4, Aug. 2021, doi: 10.1016/j.chbr.2021.100130.
[4] I. Gonta and A. Bulgac, "The adaptation of students to the academic environment in university," *Revista Romaneasca pentru Educatie Multidimensionala*, vol. 11, no. 3, pp. 34–44, Sep. 2019, doi: 10.18662/rrem/137.
[5] M. Al-Hail, M. F. Zguir, and M. Koç, "University students' and educators' perceptions on the use of digital and social media platforms: a sentiment analysis and a multi-country review," *iScience*, vol. 26, no. 8, Aug. 2023, doi: 10.1016/j.isci.2023.107322.
[6] J. Almalki, "A machine learning-based approach for sentiment analysis on distance learning from Arabic Tweets," *PeerJ Computer Science*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.1047.
[7] Z. Kastrati, F. Dalipi, A. S. Imran, K. P. Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study," *Applied Sciences*, vol. 11, no. 9, 2021, doi: 10.3390/app11093986.
[8] S. Ulfa, R. Bringula, C. Kurniawan, and M. Fadhli, "Student feedback on online learning by using sentiment analysis: a literature review," in *2020 6th International Conference on Education and Technology (ICET)*, 2020, pp. 53–58, doi: 10.1109/ICET51153.2020.9276578.
[9] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon-based approaches," in *International Conference on Research and Innovation in Information Systems (ICRIIS)*, 2017, doi: 10.1109/ICRIIS.2017.8002475.
[10] A. Q. Al-Bayati, A. S. Al-Araji, and S. H. Ameen, "Arabic sentiment analysis (ASA) using deep learning approach," *Journal of Engineering*, vol. 26, no. 6, pp. 85–93, 2020, doi: 10.31026/j.eng.2020.06.07.
[11] L. D. C. S. Subhashini, Y. Li, J. Zhang, and A. S. Atukorale, "Integration of fuzzy and deep learning in three-way decisions," in *IEEE International Conference on Data Mining Workshops, ICDMW*, 2020, pp. 71–78, doi: 10.1109/ICDMW51313.2020.00019.
[12] M. Alzaid and F. Fkih, "Sentiment analysis of students' feedback on e-learning using a hybrid fuzzy model," *Applied Sciences*, vol. 13, no. 23, 2023, doi: 10.3390/app132312956.

[13] A. C. T. Mary, "Hybrid deep learning model to predict students' sentiments in higher educational institutions," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 1, pp. 46–61, 2025.
[14] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decision Analytics Journal*, vol. 11, 2024, doi: 10.1016/j.dajour.2024.100473.
[15] K. Mrhar, L. Benhiba, S. Bourekkache, and M. Abik, "A Bayesian CNN-LSTM model for sentiment analysis in massive open online courses MOOCs," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 23, pp. 216–232, 2021, doi: 10.3991/ijet.v16i23.24457.
[16] X. Li, H. Zhang, Y. Ouyang, X. Zhang, and W. Rong, "A shallow BERT-CNN model for sentiment analysis on MOOCs comments," in *TALE 2019 - 2019 IEEE International Conference on Engineering, Technology and Education*, 2019, doi: 10.1109/TALE48000.2019.9225993.
[17] J. Zheng, J. Wang, Y. Ren, and Z. Yang, "Chinese sentiment analysis of online education and internet buzzwords based on BERT," in *Journal of Physics: Conference Series*, 2020, doi: 10.1088/1742-6596/1631/1/012034.
[18] M. Jebbari *et al.*, "A novel hybrid model for sentiment analysis in MOOC forums with hybrid word and character-level neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37, no. 3, 2025, doi: 10.11591/ijeecs.v37.i3.pp1758-1771.
[19] W. B. Cavnar, J. M. Trenkle, and A. A. Mi, "N-gram-based text categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.
[20] V. Teller, "Book reviews, speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition," *Computational Linguistics*, vol. 26, no. 4, pp. 638–641, Dec. 2000, doi: 10.1162/089120100750105975.
[21] P. A. Pandian, "Performance evaluation and comparison using deep learning techniques in sentiment analysis," *Journal of Soft Computing Paradigm*, vol. 3, no. 2, pp. 123–134, 2021, doi: 10.36548/jscp.2021.2.006.
[22] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: a state-of-the-art review," *Natural Language Processing Journal*, vol. 6, 2024, doi: 10.1016/j.nlp.2024.100059.
[23] P. S. Reddy, D. R. Sri, C. S. Reddy, and S. Shaik, "Sentimental analysis using logistic regression," *International Journal of Engineering Research and Applications*, vol. 11, pp. 36–40, 2021.
[24] J. S. Cramer, "The origins of logistic regression," *SSRN Electronic Journal*, 2005, doi: 10.2139/ssrn.360300.
[25] M. Maalouf, "Logistic regression in data analysis: an overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3, pp. 281–299, 2011, doi: 10.1504/IJDATS.2011.041335.
[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[27] A. Mishra, K. Tripathi, L. Gupta, and K. P. Singh, "Long short-term memory recurrent neural network architectures for melody generation," *Advances in Intelligent Systems and Computing*, vol. 817, pp. 41–55, 2019, doi: 10.1007/978-981-13-1595-4_4.
[28] M. S. Ahammad, S. A. Sinthia, M. M. Ahmed, F. Y. Mou, M. N. A. Ikram, and M. Ghosh, "Sentiment analysis of user-generated reviews of online education applications using deep learning and transformer learning algorithms," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT),* 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10724110.
[29] H. Peng, Z. Zhang, and H. Liu, "A sentiment analysis method for teaching evaluation texts using attention mechanism combined with CNN-BLSTM model," *Scientific Programming*, 2022, doi: 10.1155/2022/8496151.
[30] V. D. Nguyen, K. Van Nguyen, and N. L. T. Nguyen, "Variants of long short-term memory for sentiment analysis on Vietnamese students' feedback corpus," in *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 306–311, doi: 10.1109/KSE.2018.8573351.

## BIOGRAPHIES OF AUTHORS

**Raja Ouadad** received the M.Sc. degree in Computer Science from the Faculty of Sciences and Techniques, Beni Mellal, Morocco, in 2022. Currently, she is a Ph.D. candidate in the Department of Mathematics and Informatics at the Polydisciplinary Faculty, Sultan Moulay Slimane University, Morocco. Her research interests include sentiment analysis, artificial intelligence, machine learning, and deep learning algorithms. She can be contacted at email: ouadadraja2@gmail.com.



**Hicham Mouncif** received the D.Sc. degree (Doctor Habilitatus D.Sc.) in Computer Science from the Faculty of Sciences and Techniques, Beni Mellal, Morocco. He has been a professor of Computer Science at the Polydisciplinary School of Beni Mellal since 2013. He is also a director of Graduate Studies in Computer Science and head of the master of Informatics Systems Engineering (2019-present). His research interests are in e-learning, machine learning, and GIS technology applications. He was the head of the Department of Computer Sciences at the Faculty Polydisciplinaire from 2017 to 2020. He can be contacted at email: h.mouncif@usms.ma.