A Hybrid CNN-LSTM Model for Opinion Mining and Classification of Course Reviews

Hatem Majouri[a], Olfa Gaddour[b] and Yessine Hadj Kacem[c]
CES Laboratory, National Engineering, School of Sfax, University of Sfax, Tunisia

Keywords: Deep-Learning, e-Learning Platforms, CNN-LSTM, Online Course Reviews, User Feedback Analysis.

Abstract: Automatic analysis of online course reviews is a critical task that has garnered significant interest, particularly for improving the quality of e-learning platforms. The challenge lies in accurately classifying user feedback in order to generate actionable insights for educators and learners. In this work, we investigate the effectiveness of a hybrid CNN-LSTM model compared to several state-of-the-art deep learning models, including BERT, LSTM, GRU, and CNN, for analyzing reviews collected from the FutureLearn platform. Our experiments demonstrate that the proposed model achieves superior performance in classifying user reviews, with an accuracy of 0.95. These results highlight the potential of advanced deep learning techniques in extracting meaningful insights from user feedback, offering valuable guidance for course developers and learners.

1 INTRODUCTION

The rise of online education has transformed the landscape of learning, offering unprecedented access to courses across a wide range of subjects. E-learning platforms such as FutureLearn have gained popularity by enabling learners to acquire new skills and knowledge from the comfort of their homes. However, with the increasing number of courses available, evaluating and improving the quality of online education has become crucial. User reviews provide valuable insights into learner satisfaction, course effectiveness, and areas for improvement. Effective analysis of these reviews is vital for enhancing the learning experience and helping both learners and course providers make informed decisions.

Review analysis—the process of automatically detecting and classifying opinions expressed in text—has emerged as a powerful tool for understanding user feedback. Traditional methods of review analysis often struggle with the nuances of language, making it challenging to accurately capture user opinions. However, recent advances in deep learning have led to the development of sophisticated models capable of addressing these limitations. Notable models, such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), LSTM (Long Short-Term Memory) (Siami-Namini et al., 2019), GRU (Gated Recurrent Unit) (She and Jia, 2021), and CNN (Convolutional Neural Network) (LeCun et al., 1998), have shown great promise in text classification tasks, enabling more accurate review analysis.

In this paper, we explore the application of state-of-the-art deep learning techniques to the task of review analysis on a dataset of online course reviews from the FutureLearn platform. The dataset was collected using web scraping and structured to include features such as course names, student reviews, and ratings. Our study is distinctive in its comprehensive evaluation of multiple deep learning models, including BERT, LSTM, GRU, and CNN, for classifying user reviews in the context of e-learning. By leveraging these advanced techniques, we aim to uncover deeper insights into user feedback, offering a more nuanced understanding of learner experiences. This work not only contributes to the growing body of research on feedback analysis in online education but also provides practical implications for enhancing the quality of e-learning platforms.

The remainder of this paper is structured as follows. Section 2 reviews related work on review analysis and the application of deep learning techniques in the context of e-learning. Section 3 details the methodology used for data collection, including the web scraping process, dataset structuring, and de-

[a] https://orcid.org/0009-0002-6629-8527
[b] https://orcid.org/0000-0002-2693-2055
[c] https://orcid.org/0000-0002-5757-6516

Majouri, H., Gaddour, O. and Kacem, Y. H.
A Hybrid CNN-LSTM Model for Opinion Mining and Classification of Course Reviews.
DOI: 10.5220/0013177900003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 734-741
ISBN: 978-989-758-737-5; ISSN: 2184-433X

scriptions of the deep learning models implemented for feedback classification. Section 4 presents the experimental results, highlighting the performance of each model. Finally, Section 5 concludes the paper with a summary of key contributions and suggestions for future research.

2 RELATED WORK

The rise of e-learning platforms has spurred research into user feedback analysis to enhance course quality. This section examines classification methods for student reviews, focusing on datasets and methodologies to identify trends and outline future research directions.

In (Onan, 2020), deep-learning techniques were used to analyze Student Evaluations of Teaching (SET) for assessing teaching effectiveness and guiding administrative decisions. The study employed a recurrent neural network (RNN) enhanced with an attention mechanism and GloVe word embeddings, showcasing the model's capabilities. However, the dataset, not designed for educational or sentiment analysis, led to overfitting due to its small size relative to feature complexity.

In (Kastrati et al., 2020b), an aspect-based opinion mining model improved feedback analysis in online courses. Despite its potential, reliance on a single dataset raised scalability concerns, and the model's architecture failed to capture semantic relationships effectively due to limited CNN dimensions and simplistic layers.

The authors in (Chakravarthy et al., 2021) emphasized the importance of qualitative feedback in online education, which is often overshadowed by quantitative data. They developed an opinion-mining framework using NLP and machine learning to classify student feedback from a Coursera course. However, their findings were constrained by the dataset, and automated methods overlooked nuanced opinions.

In (Onan, 2021), advanced machine-learning techniques, including ensemble learning and deep-learning, were applied to MOOC reviews. The study evaluated text representation and word-embedding schemes on a dataset of 66,000 reviews but faced challenges with model interpretability and limited generalizability due to its reliance on a large dataset.

Research in (Mrhar et al., 2021) compared CNN, LSTM, and CNN-LSTM models for sentiment analysis in MOOCs. A key limitation was the reliance on manually labeled datasets, which introduced subjectivity and scalability issues for larger or more diverse datasets.

In (Koufakou, 2023), deep-learning models like CNN, BERT, RoBERTa, and XLNet were used for sentiment analysis and topic classification of student feedback. While the study offered insights into model optimization, it lacked exploration of model interpretability and employed unbalanced datasets, skewing results and impairing performance on smaller categories.

Other works, such as (El-Halees, 2011; Cabada et al., 2018; Kastrati et al., 2020a; Yan et al., 2021; Edalati et al., 2022; Shaik et al., 2023), illustrate the effectiveness of machine learning and deep-learning in course feedback analysis. Despite their contributions to enhancing the educational experience, these studies often face challenges in scalability, model interpretability, and adaptability across platforms.

3 PROPOSED APPROACH

This section presents the proposed approach, detailing a reliable system for data collection and model generalization across diverse online courses and disciplines. It describes the dataset size, preprocessing techniques, and data augmentation to address class imbalance. Dropout and L2 regularization are applied to mitigate overfitting. The hybrid CNN-LSTM model leverages CNN to capture local features and context, while LSTM preserves sequential relationships, enabling comprehensive data analysis. Performance is evaluated using metrics such as accuracy, recall, and AUC, providing a robust assessment. This approach represents a significant improvement over previous methods by combining advanced deep learning techniques with rigorous evaluation processes.

3.1 Proposed System Architecture

This subsection describes the proposed sentiment analysis architecture for e-learning platform reviews. As illustrated in Figure 1, the process starts with raw data collection and cleaning to ensure quality. The text is then tokenized or vectorized for model training, with the data split into training and testing sets for evaluation. Optimized deep-learning models are used to classify the reviews, improving accuracy and reliability in sentiment detection.

3.2 Data Collection

Data collection includes key steps such as identifying sources, web scraping, structuring the dataset, and determining its size, all essential for ensuring data accuracy and usability.

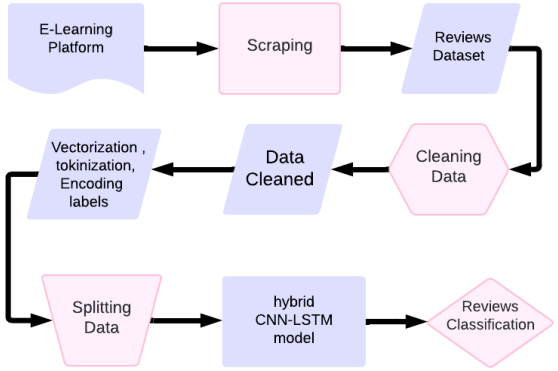

Figure 1: Overview of the Data Processing and Classification Pipeline.

3.2.1 Data Source

The data was collected from the FutureLearn platform (MAGNONI and PLUTINO, 2018), an online learning environment offering a wide range of courses from universities and institutions worldwide. FutureLearn enables students to engage in courses featuring interactive elements such as quizzes and discussions. Our focus was on compiling and analyzing student evaluations of various courses, extracting feedback on experiences, content, instructor quality, and overall satisfaction. The objective was to gain insights into the quality and effectiveness of the courses offered by FutureLearn.

3.2.2 Web-Scraping

We used web scraping to collect data for our online course review dataset, efficiently extracting reviews and insights from various platforms (Khder, 2021). Python scripts were employed to analyze website structures, with the Beautiful Soup library parsing HTML and identifying relevant tags for extraction. The gathered data was stored in CSV or XLSX files, creating a comprehensive dataset of user reviews and ratings.

3.2.3 Data Structure

The structure of a data set varies depending on the type of data and the intended use, usually consisting of rows and columns where each row represents an individual data point and each column denotes a specific attribute. In this context, the dataset captures information about online courses, student reviews, and corresponding ratings, facilitating effective data processing and analysis. Key features include course name as a string, reviews as a string, and rate as a floating, supporting comprehensive analysis of course feedback and ratings.

Table 1: Dataset Size by Class (Original vs. After Cleaning).

Class	Original Number of Samples	Number of Samples After Cleaning
1	19,271	18,978
0	7,932	5,228
Total	**27,203**	**24,206**

3.3 Data Labeling

Sentiment classification can be categorized as either binary, which involves the classification of reviews into positive or negative categories, or multi-class, which encompasses labels such as strong positive, positive, neutral, negative, and strong negative. The application of binary classification is prevalent in the field of sentiment analysis research (Tripathy et al., 2016). Furthermore, in (Guru and Bajnaid, 2023), the dataset underwent a relabeling process to facilitate binary sentiment classification through the utilization of TextBlob. Reviews exhibiting a positive polarity were designated as positive, whereas those demonstrating zero or negative polarity were classified as negative. This methodological simplification aimed to enhance the differentiation between positive and negative sentiments. The present study employed web-scraping techniques to amass a total of 30,121 reviews regarding online courses, which were initially classified into five distinct categories according to a rating scale ranging from 1.0 to 5.0. Due to a pronounced deficiency of data within classes 1, 2, and 3, our analysis concentrated on binary classification by designating 1 and 2-star evaluations as negative (0) and 3, 4, and 5-star evaluations as positive (1). The transformation of a multi-class problem into a binary classification framework is a prevalent technique in sentiment analysis, which streamlines the task to emphasize the dichotomy of positive versus negative sentiments. As highlighted in (Pang and Lee, 2008), this methodology adeptly captures critical differentiations in opinionated text and has the potential to yield more resilient models, particularly in contexts characterized by imbalanced or sparse datasets. Following the reclassification process, we discarded empty rows and duplicates, thereby enhancing the dataset to accurately reflect the distribution of each binary category. The conclusive specifications of the dataset are presented in TABLE 1.

3.4 Data Preprocessing

The initial phase in the analysis of reviews involves the meticulous preparation of textual data through

processes of cleansing and refinement. A significant proportion of the unprocessed reviews obtained from FutureLearn exhibit extraneous elements and absent information, thereby necessitating a thorough preprocessing to facilitate effective analytical procedures; to partition the dataset into training and testing subsets, an allocation of 80% is designated for training purposes while 20% is reserved for testing.

3.4.1 Data Cleaning

Data cleansing is a critical process for enhancing the quality of datasets by rectifying inconsistencies, errors, and absent values. Fundamental activities encompass label encoding to transform target variables into binary classifications, addressing missing and duplicate entries, and ensuring uniformity in data types. Text preprocessing procedures involve the elimination of special characters, HTML elements, and extraneous spaces, the expansion of contractions, tokenization, lowercasing, and lemmatization of lexemes. Stopwords are excluded, while significant negations such as "no" and "not" are preserved in order to maintain the contextual integrity of sentiment. These methodologies guarantee that the data is adequately prepared for the training of models.

3.4.2 Data Transformation

Data transformation for text classification varies by model. LSTM, GRU, and CNN use tokenization (splitting text into words or subwords), while BERT relies on its specialized subword tokenizer. For LSTM, GRU, and CNN, text normalization (e.g., lowercasing and punctuation removal) is applied, whereas BERT handles this internally via its tokenizer. Padding and truncation ensure consistent sequence lengths, performed manually for LSTM, GRU, and CNN, but automated by BERT. Tokens are converted into numerical formats using embeddings or one-hot encoding for LSTM, GRU, and CNN, and contextual embeddings for BERT. Categorical labels are encoded numerically across all models.

3.5 CNN-LSTM Model Architecture

Our proposed hybrid CNN-LSTM architecture, shown in Figure 2, is designed for text data analysis and classification. Input sentences are represented as a matrix of size $N \times K$, where N is the number of sentences and K the number of features (e.g., word embeddings). A convolutional layer processes this matrix, applying filters to extract local features and capture essential patterns and context.

The convolutional output is passed through a max-pooling layer, which reduces dimensionality by selecting maximum values from each region, retaining critical features while reducing computational cost. The pooled output is then fed into an LSTM layer, which captures dependencies and contextual information across sequences.

Subsequently, the features are processed by one or more fully connected (dense) layers for high-level reasoning. A dropout layer follows to prevent overfitting by randomly deactivating neurons during training, enhancing generalization. The final output is generated through a sigmoid activation function, producing a probability between 0 and 1, representing the likelihood of each class. This output classifies sentences into two categories: *Class 0* and *Class 1*.

By combining convolutional layers for feature extraction with LSTM layers for sequence processing, followed by dense and dropout layers, this architecture achieves robust classification performance.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Evaluation Metrics

To evaluate our algorithm's performance, we use several metrics. Accuracy, calculated as the proportion of correctly classified instances among all instances, is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Precision indicates the proportion of true positive predictions among all positive predictions made by the model and is given by:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall represents the proportion of true positive predictions among all actual positive instances and is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

The F1 Score, which is the harmonic mean of precision and recall, provides a single metric that balances both aspects and is defined as:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

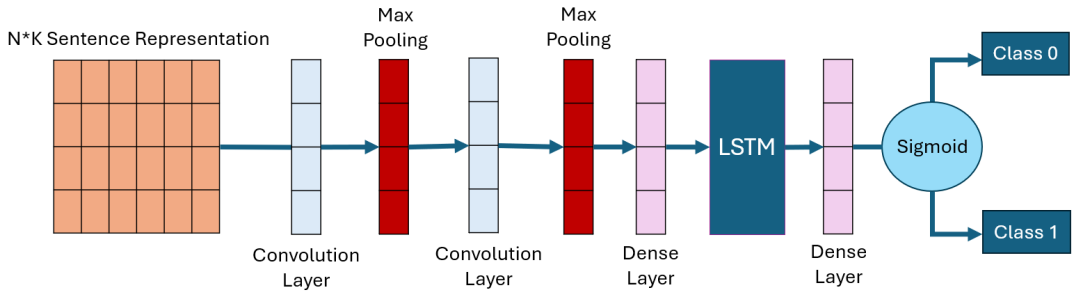

Figure 2: Proposed Hybrid CNN-LSTM Model Architecture.

4.2 Performance Visualization

Key visualizations for assessing model performance include the accuracy curve, which tracks changes in accuracy over epochs, and the loss curve, which highlights error reduction and convergence. The confusion matrix displays the distribution of true positives, false positives, and false negatives across classes, while the ROC curve demonstrates the trade-off between true and false positive rates at various thresholds.

4.3 Experimental Scenario

The dataset was scraped, cleaned, tokenized, and split into 80% training and 20% testing, with augmentation applied only to the training data. Tokenized words were vectorized using GloVe embeddings to capture semantic associations. The model combines CNN and LSTM layers for feature extraction and sequence analysis. A non-trainable GloVe embedding layer was followed by CNN layers with kernels, and finally, an LSTM layer with regularization to mitigate overfitting. Dropout layers further enhanced generalization, while dense layers integrated features for classification. The model was trained using binary cross-entropy loss and the Adam optimizer for 50 epochs with a batch size of 32, tracking performance on the test set.

The CNN architecture includes two Conv1D layers with 128 filters and a kernel size of 5, each followed by MaxPooling1D layers with a pool size of 2. An additional Conv1D layer with 64 filters and a kernel size of 3, followed by another MaxPooling1D layer, is included. All CNN layers use dropout with a rate of 0.3%. The dense layers after the CNN have 32 and 16 units with ReLU activation and L2 regularization applied to both kernel and bias. The LSTM layer matches the embedding dimension, with $return_sequences = False$ to output the hidden state. L2 regularization and a 0.3 dropout rate are also applied to the LSTM layer.

4.4 Results Without Data Augmentation

In the results section, we analyze the raw data processed without augmentation, evaluating the performance of GRU, LSTM, CNN, and BERT models. Each model is assessed to highlight its ability to handle the original data, along with its relative advantages and limitations.

TABLE 2 presents the performance results of these models applied to our dataset without any augmentation.

Starting with BERT, the model shows a precision of 56% for class 0 and 94% for class 1. The recall values are 82% for both classes, and the F1 scores are 67% for class 0 and 88% for class 1. The overall accuracy of BERT is 82

Next, the LSTM model achieves a precision of 79% for class 0 and 88% for class 1. The recall for class 0 is 52% and 96% for class 1. The F1 scores are 62% for class 0 and 92% for class 1, with an overall accuracy of 87%.

The GRU model provides a precision of 82% for class 0 and 92% for class 1. The recall is 71% for class 0 and 95% for class 1. The F1 scores are 76% for class 0 and 94% for class 1, resulting in an overall accuracy of 90%.

The CNN model reports a precision of 79% for class 0 and 94% for class 1. The recall values are 79% for class 0 and 94% for class 1, and the F1 scores are 79% for class 0 and 94% for class 1, with an overall accuracy of 91%.

In the absence of data augmentation, the CNN-LSTM model performs well with an accuracy of 92%. Precision is 90% for class 0 and 94% for class 1, demonstrating high accuracy in detecting positive cases. The recall for class 0 is 85%, while for class 1 it is 92%, indicating effective capture of genuine positives. The F1 scores are 87% for class 0 and 93% for class 1, reflecting a strong balance between precision and recall.

Table 2: Performance metrics without and with augmentation.

Performance	Before Augmentation					After Augmentation				
	BERT	LSTM	GRU	CNN	**CNN-LSTM**	BERT	LSTM	GRU	CNN	**CNN-LSTM**
Precision	86.5%	87%	83.5%	75%	**92%**	77%	87%	88%	87.5%	**93%**
Recall	86.5%	83%	74%	82%	**88.5%**	81.5%	84.5%	87.5%	87.5%	**92.5%**
F1 Score	86.5%	80%	77%	77.5%	**91.5%**	78.5%	86%	88%	88%	**92.5%**
Accuracy	82%	87%	90%	91%	**92%**	84%	91%	92%	92%	**95%**

4.5 Results with Data Augmentation

Data augmentation enhances model performance by improving precision, recall, and F1 scores, increasing accuracy and generalization. TABLE 2 shows our comparison results after data augmentation.

Starting with BERT, data augmentation results in a precision of 61% for class 0 and 93% for class 1. The recall values are 76% for class 0 and 87% for class 1, with F1 scores of 67% for class 0 and 90% for class 1. The overall accuracy of the model improves to 84%.

For LSTM, precision reaches 81% for class 0 and 93% for class 1. Recall values are 74% for class 0 and 95% for class 1, while the F1 scores are 78% for class 0 and 94% for class 1. The total accuracy of the model is 91%.

GRU shows a precision of 81% for class 0 and 95% for class 1, with recall values of 80% for class 0 and 95% for class 1. The F1 scores are 81% for class 0 and 95% for class 1, resulting in an overall accuracy of 92%.

The CNN model achieves a precision of 80% for class 0 and 95% for class 1. Recall values are 81% for class 0 and 94% for class 1, with F1 scores of 81% for class 0 and 95% for class 1. The overall accuracy is 92%.

With data augmentation, the CNN-LSTM model shows a significant performance boost. Precision rises to 89% for class 0 and 97% for class 1. Recall improves to 90% for class 0 and 95% for class 1, indicating enhanced detection of true positives. The F1 scores increase to 89% for class 0 and 96% for class 1. Overall accuracy surpasses 95%, reflecting a substantial improvement in model performance due to the augmentation strategies.

4.6 Performance Visualization

In the visualization section, we evaluate the accuracy and loss curves of the hybrid LSTM-CNN model, analyze the confusion matrix, and plot the ROC curve to assess class distinction.

4.6.1 Accuracy Curve

The accuracy curve in Figure 3 illustrates the model's performance over 50 training epochs for both the training and validation datasets. Initially, the training accuracy increases rapidly, reaching near-perfect levels around the 10^{th} epoch, where it then plateaus. The validation accuracy also rises quickly during the early epochs, stabilizing around 90-92% after the 10^{th} epoch. The gap between training and validation accuracy suggests minor overfitting, indicating the model performs well on training data but generalizes less effectively to new data.

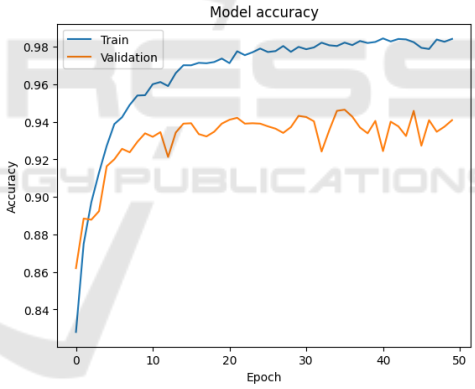

Figure 3: Accuracy Curve.

4.6.2 Loss Curve

The loss curve in Figure 4 represents the model that initially learns effectively, with both training and validation losses decreasing rapidly. However, after about 10 epochs, the validation loss plateaus while the training loss continues to decrease, suggesting that the model may be overfitting to the training data. This divergence implies that the model is becoming too specialized for the training set, potentially compromising its ability to generalize well to new, unseen data.

4.6.3 Confusion Matrix

The confusion matrix depicted in Figure 5 summarizes 1,046 instances of Class 0, it correctly predicted

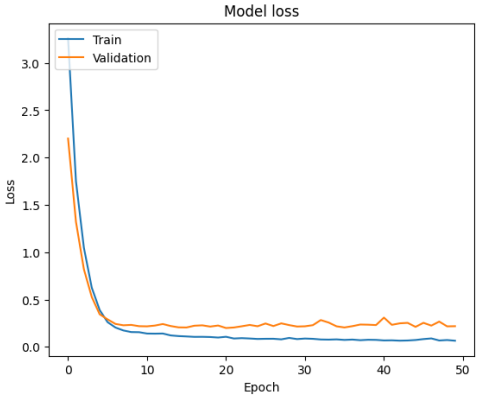

Figure 4: Loss Curve.

856 and misclassified 190 as Class 1. For Class 1, out of 3,796 cases, it accurately predicted 3,711 but misclassified 190 as Class 0. While the model shows a high number of true positives and true negatives, the presence of false positives (85) and false negatives (190) indicates areas for improvement in its predictive capability.

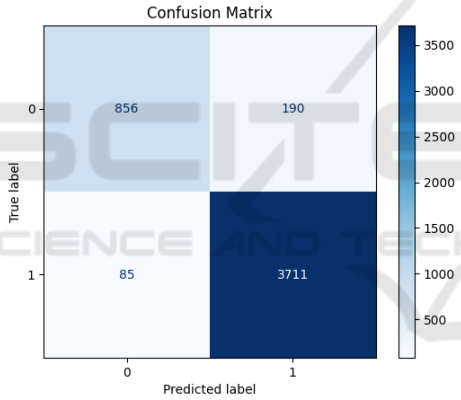

Figure 5: Confusion Matrix.

4.6.4 ROC Curve

The ROC curve shown in Figure 6 model's ROC curve is close to the top-left corner, indicating a high level of classification performance. The area under the curve (AUC) is 0.95, which suggests that the model has excellent discrimination ability. A higher AUC value closer to 1.0 implies that the model is very effective at distinguishing between the positive and negative classes, with minimal overlap.

4.7 Discussion

The results in the curves and tables underscore the effectiveness of different models in binary text classification tasks, with particular emphasis on the hybrid CNN-LSTM model. The accuracy and loss curves in-

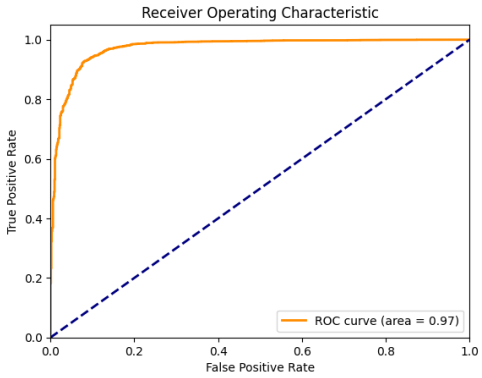

Figure 6: Roc Curve.

dicate strong learning capabilities, with high training and consistently robust validation accuracy, suggesting minimal overfitting. This is corroborated by the confusion matrix, which shows the model's adeptness at distinguishing between classes, achieving high precision and recall.

The ROC curve reinforces the model's performance, with a high area under the curve (AUC), indicating strong discrimination between positive and negative classes. Comparative analysis in TABLE 2 reveals that while BERT and CNN models perform well, particularly in Class 1, the CNN-LSTM model provides the best balance of precision, recall, and F1 score, making it the top performer.

The augmented results in TABLE 2 further highlight the CNN-LSTM model's superiority, achieving 95% accuracy while maintaining excellent precision and recall. The improvement with data augmentation is notable in the LSTM and GRU models, which show significant gains in recall. The CNN-LSTM model remains the most robust, with the highest overall accuracy and balanced metrics, underscoring its suitability for complex classification tasks.

5 CONCLUSIONS

In conclusion, this paper presents a comprehensive study on the automated classification of student reviews in e-learning platforms using a large dataset from Future Learn. The primary contribution is the development of a hybrid deep learning architecture that combines convolutional neural networks (CNN) and long-short-term memory (LSTM) networks. This approach achieves remarkable performance, with an accuracy of 95%, highlighting the robustness of deep learning in processing and classifying large-scale educational data. The dataset's size and diversity, coupled with the model's capabilities, underscore its relevance in advancing opinion analysis for online ed-

ucation and improving e-learning platforms through sophisticated AI techniques.

Looking ahead, this research lays the groundwork for enhancing model interpretability and expanding hybrid architectures to broader educational data mining contexts. Further improvements could include exploring methods to boost model performance and scalability. Expanding the dataset may lead to deeper insights into student feedback, driving enhancements in e-learning platform effectiveness and user satisfaction.

REFERENCES

Cabada, R. Z., Estrada, M. L. B., and Bustillos, R. O. (2018). Mining of educational opinions with deep learning. *Journal of Universal Computer Science*, 24(11):1604–1626.

Chakravarthy, V., Kameswari, M., Mydeen, H. D., and Seenivasan, M. (2021). Opinion mining from student text review for choosing better online courses. In *IOP Conference Series: Materials Science and Engineering*, page 012067. IOP Publishing.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Edalati, M., Imran, A. S., Kastrati, Z., and Daudpota, S. M. (2022). The potential of machine learning algorithms for sentiment classification of students' feedback on mooc. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*, pages 11–22. Springer.

El-Halees, A. (2011). Mining opinions in user-generated contents to improve course evaluation. In *Software Engineering and Computer Systems: Second International Conference, ICSECS 2011, Kuantan, Pahang, Malaysia, June 27-29, 2011, Proceedings, Part II 2*, pages 107–115. Springer.

Guru, C. and Bajnaid, W. (2023). Prediction of customer sentiment based on online reviews using machine learning algorithms. *International Journal of Data Science and Advanced Analytics*.

Kastrati, Z., Arifaj, B., Lubishtani, A., Gashi, F., and Nishliu, E. (2020a). Aspect-based opinion mining of students' reviews on online courses. In *Proceedings of the 2020 6th International conference on computing and artificial intelligence*, pages 510–514.

Kastrati, Z., Arifaj, B., Lubishtani, A., Gashi, F., and Nishliu, E. (2020b). Aspect-based opinion mining of students' reviews on online courses. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, ICCAI '20, pages 510–514. ACM.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).

Koufakou, A. (2023). Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies*, 29(3):2973–2997.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

MAGNONI, F. and PLUTINO, A. (2018). The moveme project on the futurelearn platform. some considerations after the first pilot. *E-Learning, MOOC e Lingue Straniere: Studi, Ricerche e Sperimentazioni E-Learning, MOOCs and Foreign Languages: Research, Studies and Experiences*, page 103.

Mrhar, K., Benhiba, L., Bourekkache, S., and Abik, M. (2021). A bayesian cnn-lstm model for sentiment analysis in massive open online courses moocs. *International Journal of Emerging Technologies in Learning (iJET)*, 16(23):216–232.

Onan, A. (2020). Mining opinions from instructor evaluation reviews: a deep learning approach. *Computer Applications in Engineering Education*, 28(1):117–138.

Onan, A. (2021). Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Computer Applications in Engineering Education*, 29(3):572–589.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., and Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003.

She, D. and Jia, M. (2021). A bigru method for remaining useful life prediction of machinery. *Measurement*, 167:108277.

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*, pages 3285–3292. IEEE.

Tripathy, A., Agrawal, A., and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126.

Yan, X., Li, G., Li, Q., Chen, J., Chen, W., and Xia, F. (2021). Sentiment analysis on massive open online course evaluation. In *2021 International Conference on Neuromorphic Computing (ICNC)*, pages 245–249. IEEE.