ISNNI: 3007-3391

HEC Category: Y

Journal of Computers and Intelligent Systems

Vol. 4, No. 1, 2023, pages 22 - 30

Journal Homepage:

https://journals.iub.edu.pk/index.php/JCIS/

Harnessing Encoder-based Transformer Models for Multilingual Student Feedback Sentiment Analysis

Majdah Alvi[1], Muhammad Ali Qureshi[2], Muhammad Bux Alvi[1],

[1]Department of Computer Systems Engineering, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan

[2]Department of Information and Communication Engineering, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan

ARTICLE INFO

Article History:

Received:	March	01, 2026
Revised:	March	15, 2026
Accepted:	March	25, 2026
Available Online:	March	31, 2026

Keywords:

Encoder-based Transformer Models
Large Language Models
Multilingual Sentiment Analysis
Student Feedback
Textual Data

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Encoder-based Transformer models (ETMs) have demonstrated outstanding performance in sentiment analysis, comparable to human capability. ETMs can be applied to understand the students' perspectives that they provide through textual feedback using semi-automated feedback collection methods. The semi-automatic feedback collection includes online social networks, University blogs, personal interviews, and Google forms. However, analyzing multilingual student feedback is challenging, especially when it includes code-switched expressions. The difficulty becomes twofold when the feedback contains resource-poor language scripts. Traditionally, student feedback was processed using a rule-based approach or machine learning algorithms (MLAs), which proved inadequate. Deep neural networks and their variants demonstrated good results but required labeled training data. Encoder-based transformer pre-trained models have shown promising results. However, their performance in multilingual datasets (English, Urdu, and Roman Urdu) remains suboptimal. This study collected a multilingual dataset comprising 11,686 student feedback samples and explored the potential of encoder-based transformer models in automating multilingual sentiment analysis using student feedback. The dataset comprises of 33.2% English, 33.3% Urdu, and 33.5% Roman Urdu student feedback expressions. The applied ETM model identified 5,730 sentences as positive, 4,232 as negative, and 1,724 as neutral, achieving enhanced performance compared to the generic ETM. Additionally, it identifies challenges and limitations in multilingual feedback analysis and offers recommendations for model selection and fine-tuning.

Corresponding Author's Email: majida.alvi@iub.edu.pk

1. Introduction

Large language models (LLMs) are built on the Transformer architecture. Transformer architectures are primarily encoder-based, decoder-only, or encoder-decoder models. In all cases, LLMs demonstrate excellent performance in natural language processing (NLP) tasks, including text classification, question answering, and sentiment analysis. Sentiment analysis is an important task in NLP that determines the polarity (positive, negative, or neutral) of an opinionated text [1]. Today, sentiment analysis systems have advanced to the point where their accuracy in automatically identifying polarity is now comparable to that of human capability. Therefore, it is the ideal time to capitalize on technological advances to enhance our understanding of student textual feedback in higher education. The application of sentiment analysis has increased manifold as massive open online courses (MOOCs) have gained widespread acceptance alongside traditional face-to-face education. Students are unlikely to choose a new program or institution without reading previous student reviews. Manual processing of unilingual students' feedback is a tedious task due to its volume, and analyzing multilingual student comments makes it more challenging. Furthermore, it is inadequate to perform student feedback analysis using rule-based approaches or traditional machine learning methods.

With the rebirth of neural networks, their newer representation, deep learning (DL), and the availability of vast amounts of text data, the feasibility of sentiment analysis became possible. Academicians and researchers have begun performing sentiment analysis using various variants of deep neural networks, including recurrent neural networks (RNNs), long short-term memory (LSTM), and gated recurrent units (GRUs). All of these architectures have shown adequate sentiment classification results, although RNN, LSTM, and GRU have some inherent demerits [2]. These models, based on different deep learning architectures, are expected to perform well when trained on large amounts of data; that is, the larger the training data, the better the performance. The stated problem remains the bottleneck: the availability of a large amount of labeled data for training the model.

Recent breakthroughs in the form of LLMs (transformer-based, pre-trained models) have revolutionized the world. Not many AI advances, aside from ChatGPT and similar tools, have achieved such widespread popularity outside the computing community. The ability of LLMs to process text is approaching human capacity; more importantly, it does not require a training phase, eliminating the need for a training set at the user's end. This study explores the potential of encoder-based transformer models in automating multilingual sentiment analysis using student feedback. The following are important contributions to this paper:

- Preparation of a multilingual student feedback textual dataset.
- Harnessing encoder-based transformer models for sentiment analysis on the collected dataset.
- Comparison of the models developed to evaluate their suitability for multilingual student feedback sentiment analysis.
- Exploration of potential improvements and proposal of suitable solutions

The limitations of encoder-based models for multilingual sentiment analysis include: curse of multilingualism [3], cultural and linguistic differences [4], and data imbalance [5]. XLM-R is recommended for multilingual sentiment analysis [6], but fine-tuning is required for domain-specific adaptability [7].

The paper is organized as follows. Section 1 introduces the study. The state of the art is presented in Section 2, followed by a description of the development process in Section 3. Section 4 presents the results using graphs and tables. Finally Section V includes the conclusion along with limitations of the available resources.

2. Related Work

Sentiment analysis is an NLP task that focuses on determining the polarity (positive, negative, or neutral) of subjective, free-text data [8]. This section presents a historical overview of sentiment analysis, followed by recent advances in the field and related works.

2.1 From NLP to Sentiment Analysis

Sentiment analysis is primarily an NLP task. In the 1950s and beyond, NLP tasks were considered to be handled and understood through sentence structure and grammar, opening the way for rule-based approaches to accomplish them. Rule-based approaches were effective with smaller datasets. With the availability of datasets in the 1990s and early 2010s, statistical models began to emerge. N-grams [9] were adopted to capture a partial context in the text data. The next milestone was achieved through supervised machine learning algorithms [10]. The rebirth of neural networks (2010s) was a redefining moment in sentiment analysis after the vanishing gradient issue was resolved. Neural networks, increasing computational power, and access to textual data from online social networks for training started opening new avenues. The RNNs [11] and their variants are gaining acceptance. The LSTM and GRU [12] can better comprehend the context of text data. Better neural network architectures and improved pre-trained static word embeddings (Word2Vec [13], GloVe [14], FastText [15]) have brought commercial feasibility and academic acceptability to sentiment analysis and NLP.

Vaswani et al. [16] revolutionized NLP by introducing an attention mechanism, paving the way for the transformer architecture [17]. Based on query, key, and value mechanisms, the transformer architecture performs language translation through seq2seq modeling, its original task (language translation). However, the transformer architecture has also shown great potential for classification, topic extraction, and sentiment analysis, outperforming deep learning models. The encoder-based BERT (Google) [16] and GPT models by OpenAI use transformer architectures. The trends in literature on sentiment analysis primarily focus on different encoder-based transformer models.

2.2 Student Feedback Sentiment Analysis

Sentiment analysis is widely used to analyze students' perspectives. Here is an overview of recent research on the topic. The authors in [17] investigated the potential of LLMs for analyzing multiclass sentiment in student feedback about teachers. They collected the dataset and sought expert annotators to label it for sentiment polarity. The authors claimed to achieve an impressive F1-score of 88%, outperforming state-of-the-art transformer-based and deep learning models.

Edalati et al. in [18] evaluated various machine learning and deep learning models for aspect-based opinion mining on a manually annotated large-scale students' comments collected from Coursera. Their work was focused only on the teaching-related aspects of MOOCs. The authors claimed to achieve promising results of 98.01% and 99.43% F1-score using a random forest model. The two scores were computed for the topic (aspect) identification and the sentiment classification (topic-wise).

The authors of [19] suggested that human-centric student feedback analysis systems were viable for class courses involving a couple of tens of students but impractical for large-scale cases. The authors proposed a weakly supervised method for training MLA-based models on a small dataset, which effectively identified topics (aspects) in the unlabeled portion of student feedback. Consequently, it reduced the need for manual annotation, a major challenge for DL techniques. They used a dataset containing 5989 students' feedback samples (0.105 million). The experimental results indicate that the proposed framework achieved promising performance in both aspect category identification and aspect sentiment classification.

The authors in [20] encompassed a comparative study of DL models for aspect-based sentiment analysis (ABSA) on student feedback to place different approaches in context. Sindhu et al. in [21] employed an LSTM-based method to perform ABSA from the student's textual perspective. They proposed a two-stage model in which the first stage extracts polarity and the second identifies the relevant sentiment aspect.

The authors proposed a sentiment analysis system (SAS) that leveraged automatic, Global Vectors Word Embeddings (GloVe), and Bidirectional Encoder Representations from Transformers (BERT). The SAS consisted of an attention layer and a model, selected through rigorous experimental comparisons. The authors used a dataset of 3,820 student comments, including 2,773 organic comments and 1,047 generated with Generative AI. The Gen-AI was used to balance the dataset. They used RNNs, GRUs, LSTMs, and bidirectional LSTMs, with and without pre-trained GloVe embeddings. The authors claimed to achieve an F-score of 67% to 69%. In contrast, the BERT-based sentiment model implemented in Keras achieved higher F-scores (from 83% to 87%). The Bi-LSTM outperformed other models. With an attention layer, it achieved improved performance, resulting in F-scores of 89% and 88% [22].

Recent research studies focus primarily on developed (rich) languages. The least attention is paid to resourceless scripts, creating a gap. The current study aims to leverage an available encoder-based transformer model for multilingual sentiment analysis of student feedback to extract optimal insights from the collected dataset.

3. Proposed Architecture

Figure 1 shows an abstract model for multilingual sentiment analysis for student feedback. The following sub-sections elaborate on each component of the proposed model.

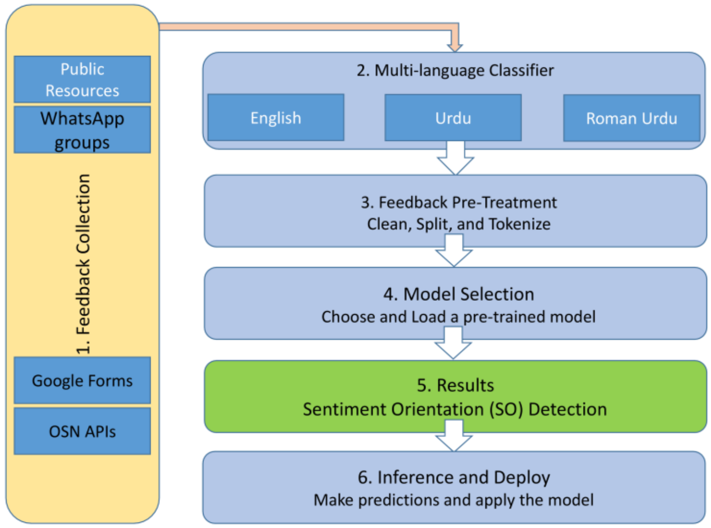

Figure 1. Multilingual Student Feedback Sentiment Analysis Model

3.1 Feedback Collection

A heterogeneous feedback method was used to collect the Multilingual Student Feedback Dataset (MLSFD) for this study. The heterogeneous approach encompassed online social networks (OSNs), WhatsApp groups, personal interviews, and Google Forms. The heterogeneous data collection approach helped ensure compliance with ethical data privacy and protection standards. The use of OSNs and Google Forms allowed students to express their perspectives freely and anonymously. The seed dataset comprises 11,686 student feedback sentences in English, Urdu, and Roman Urdu (RU). Figure 2 displays the language-wise distribution of the dataset, indicating a balanced distribution. There are 33.2% English sentences, 33.3% Urdu comments, and 33.5% Roman Urdu expressions.

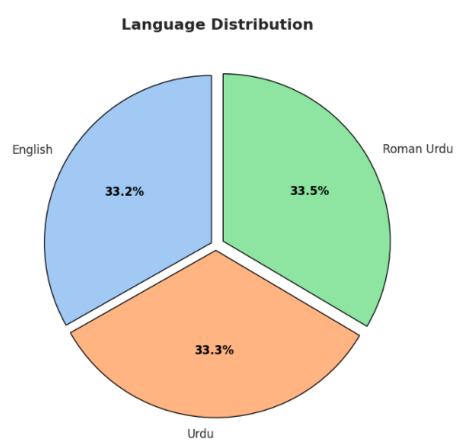

Figure 2. Language-wise distribution of MLSFD

Textual feedback was collected from various departments, primarily from students of the Faculty of Engineering and Technology at The Islamia University of Bahawalpur. Figure 3. shows a sample of ten sentences from a multilingual dataset. Three post-graduate students manually annotated the feedback into three sentiment classes (i.e., Positive, Negative, or Neutral). For each instance of the feedback, the annotators were asked to vote in one of three polarities. After annotation, only those instances were retained for which at least two annotators agreed on the same polarity.

	comments
0	Assignments are too difficult.
1	Helpful and polite teacher.
2	Classroom hamesha saaf tha.
3	کورس کا مواد پرانا ہے۔
4	Class bohot shor tha.
5	استاد موضوعات جلدی جلدی پڑھاتے ہیں۔
6	موضوعات میں وضاحت کی کمی ہے۔
7	Kabhi lectures repeat hotay hain.
8	Teacher encourages participation.
9	Teacher bohot strict hain.

Figure 3. A sample of ten sentences from MLSFD

Table 1 presents a detailed statistical analysis of MLSFD. It provides information on the number of positive, negative, and neutral sentences in each language. It also provides insight into the total number of words, the minimum and maximum word counts, and the average word length for each category across all three languages.

Table 1. MLSFD Statistics

No.	Languages	Sentences	Category	T_Sentences	T_Words	Average	Words_Min	Words_Max	**Suggestion**
1	English	3882	Positive	2,023	12,657	6.26	1	31	12/3
			Negative	1,466	9,036	6.16			
			Neutral	389	2,756	7.08			
2	Urdu	3888	Positive	2,025	15,603	7.71	1	34	
			Negative	1,468	10,885	7.41			
			Neutral	391	3,232	8.27			
3	RU	3913	Positive	2,037	15,229	7.48	1	30	
			Negative	1,481	10,639	7.18			
			Neutral	391	3,160	8.08			

3.2 Multi-language Classifier

The collected MLSFD, comprising 11,686 sentences, included English, Urdu, and Roman Urdu sentences. No single pre-trained encoder-based transformer model can manage them proficiently. Therefore, a multilingual classifier was developed to categorize the mixed dataset into three writing styles: English, Urdu, and Roman Urdu.

3.3 Feedback Pre-Treatment

Pre-treatment of textual feedback involves removing duplicate samples, removing unnecessary tokens, and tokenizing them for model compatibility. Ultimately, the entire dataset is split into training and test sets using a suitable splitting strategy. The student feedback was partially preprocessed as suggested in [23] without the recursive component. Additionally, the dataset was augmented using NLP techniques.

3.4 ETM Selection

Model selection is a crucial step for performing sentiment analysis using pre-trained encoder-based transformer models. It requires an exhaustive search to choose a pre-trained model that matches the problem type and domain. It involves searching for task-specific models, checking for model architectures (BERT, RoBERTa, DistilBERT, XML-R), model size, licensing, training domain, and pre-trained weights. The subsequent steps of polarity detection are discussed in section 4. However, the deployment step is performed using Hugging Face, Streamlit, and Gradio.

4. Results and Discussion

Figure 4 provides the sentiment polarity detection ratio for the 11,686 multilingual student feedback dataset. Out of 11,686 feedback items, the resultant multilingual ETM model identified 5,730 sentences as positive, 4,232 as negative, and 1,724 as neutral. With the generic ETM model, 5,544 sentences were detected as positive, 3,407 comments were identified as negative, and 2,735 as neutral. Here, a noticeable difference is evident in the comparison. Without a multilingual ETM, the accuracy of corrected student feedback is low. 186 more sentences were detected as positive, 825 more were identified as negative, and 1,011 neutral sentences were classified as positive or neutral. The neutral student feedback also consisted of suggestions for improving the quality of teaching, methodology, curriculum, and facilities, as shown in Table 2.

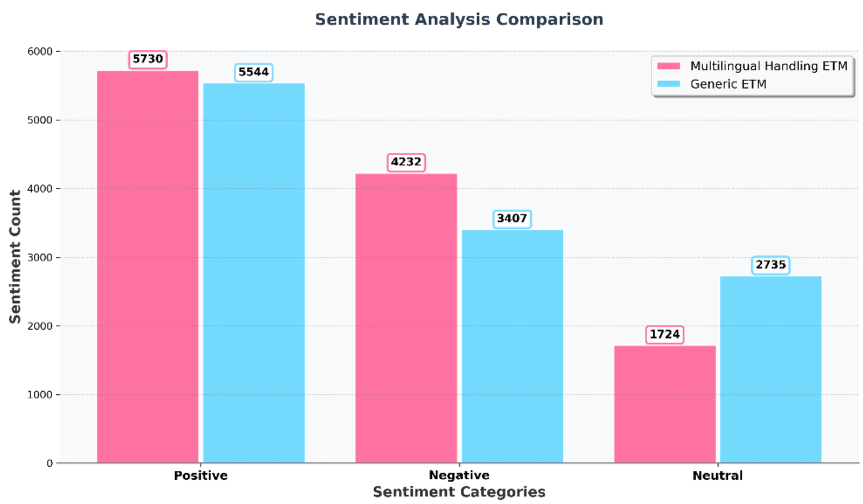

Figure 4. Multilingual Sentiment Polarity Detection using MLSFD dataset

Table 2. Student Feedback Suggestions

No.	Student Suggestions
1	syllabus me ai tools istemal shamil kerna chahye.
2	Students should be provided free access to research tools.
3	canteen ka khana sasta hona chahye
4	There should be regular counseling sessions and career guidance workshops.
5	University should facilitate free access to e-books

Table 3 presents the sentiment polarity obtained from seven multilingual encoder-based pre-trained transformer models. It presents results for the dataset sample as provided in Figure 3. In sequence 1, all the models agree to assign a "negative" polarity to the student's textual feedback, with varying degrees of confidence, except for the last two models, which employ a different polarity scheme. The confidence level for assigning sentiment polarity ranges from 0.515 to 0.989, raising questions about the robustness of the models used without human intervention. Similarly, sample two is declared "positive" with a huge difference in confidence. The data instances numbers 5 and 7 are "negative", with varying confidence. The five models did not agree on assigning a uniform sentiment polarity to samples at 3, 4, 6, and 10.

Ultimately, samples 8 and 9 were determined to be positive or biased towards positivity. Again, with a varying degree of confidence. Model_1 and Model_2 provided binary sentiment classification, whereas Models_3, 4, and 5 provided multilevel classification. Model_5 even presented five-degree results, which are very promising for gaining a better insight.

Table 3. Performance Analysis of different Multilingual encoder-based Transformer models

Model_1			Model_2		Model_3		Model_4		Model_5		nlptwon		meandmichael8011	
Light BERT			BERT		RoBERTa		XLM		RoBERTa		XLM		XLM	
No.	Sentiment Score	Confidence	Sentiment Score	Confidence	Sentiment Score	Confidence	Sentiment Score	Confidence	Sentiment Score	Confidence	Sentiment Score	Confidence	Sentiment Score	Confidence
1	Neg.	0.593	Neg.	0.989	Neg.	0.521	Neg.	0.734	Neg.	0.515	2-star	0.492	LABEL_0	0.922
2	Pos.	0.492	Pos.	0.681	Pos.	0.474	Pos.	0.502	very Pos.	0.828	5-star	0.600	LABEL_1	0.997
3	Neg.	0.498	Pos.	0.818	Neg.	0.524	Neg.	0.698	Neg.	0.447	1-star	0.379	LABEL_0	0.872
4	Pos.	0.383	Pos.	0.551	neutral	0.591	neutral	0.645	Pos.	0.346	5-star	0.432	LABEL_0	0.706
5	Neg.	0.459	Neg.	0.935	Neg.	0.471	Neg.	0.651	Neg.	0.481	1-star	0.371	LABEL_0	0.967
6	Pos.	0.307	Neg.	0.509	neutral	0.516	neutral	0.577	neutral	0.407	4-star	0.369	LABEL_1	0.934
7	Neg.	0.559	Neg.	0.816	Neg.	0.511	Neg.	0.681	Neg.	0.465	3-star	0.529	LABEL_0	0.919
8	Pos.	0.293	Pos.	0.957	neutral	0.501	neutral	0.473	very Pos.	0.607	5-star	0.280	LABEL_1	0.674
9	Pos.	0.411	Pos.	0.482	neutral	0.468	neutral	0.393	very Pos.	0.790	4-star	0.478	LABEL_1	0.910
10	Pos.	0.494	Pos.	0.994	Neg.	0.516	Neg.	0.600	Neg.	0.483	2-star	0.405	LABEL_0	0.991
Model_1: tabularisai/multilingual-sentiment-analysis Model_2: TankuVie/bert-base-multilingual-uncased-vietnamese_sentiment_analysis Model_3: MonkeyDLLLLLLuffy/CustomModel-multilingual-sentiment-analysis Model_4: MonkeyDLLLLLLuffy/CustomModel-multilingual-sentiment-analysis-enhanced Model_5: RnzB6/finetuned-multilingual-sentiment-analysis														

The last two models in Table 3 present results for two more multilingual pre-trained encoder-based transformer models. The results are presented as star rankings (nlptwon), which can be confusing, or as binary classification (meandmichael8011), which often proves inadequate for deeper analysis.

5. Conclusion

This study resulted in the development of a multilingual student feedback dataset (MLSFD) comprising 11,686 sentences in English, Urdu, and Roman Urdu. The developed encoder-based transformer model provided improved performance when compared to generic EMTs. Moreover, the paper focused on exploring the potential of pre-trained encoder-based transformer models (ETMs) for multilingual sentiment analysis, thereby eliminating the need for large-scale model training. Multilingual models suffer from inherent limitations, such as language coverage. mBERT and XLM-R cover over 100 languages; others, such as nlptown/bert-base-multilingual-uncased-sentiment, are limited to just 6 languages. Additionally, the size of the models impacts computational requirements; larger models, such as XLM-R (550M parameters), necessitate more powerful hardware. Domain adaptation is another crucial factor, as models pre-trained on general domains usually do not align well with the specific nuances of student feedback, requiring fine-tuning for optimal performance. Lastly, multilingual sentiment analysis can introduce biases due to linguistic imbalance and cultural differences, necessitating careful selection and evaluation of models. It is recommended to leverage XLM-R for model selection and fine-tuning due to its superior adaptability to diverse linguistic structures.

6. References

1. Jawad, K., et al., *Rusas: Roman urdu sentiment analysis system.* Computers, Materials, & Continua, 2024. **79**(1): p. 1463.
2. Abiodun, O.I., et al., *State-of-the-art in artificial neural network applications: A survey.* Heliyon, 2018. **4**(11).
3. Gurgurov, D., T. Bäumel, and T. Anikina, *Multilingual large language models and curse of multilinguality.* arXiv preprint arXiv:2406.10602, 2024.
4. Liu, J., et al., *Application of deep learning-based natural language processing in multilingual sentiment analysis.* Mediterranean Journal of Basic and Applied Sciences (MJBAS), 2024. **8**(2): p. 243–260.
5. Xu, Y., et al., *A survey on multilingual large language models: Corpora, alignment, and bias.* Frontiers of Computer Science, 2025. **19**(11): p. 1911362.
6. Krasitskii, M., et al., *Comparative approaches to sentiment analysis using datasets in major european and arabic languages.* arXiv preprint arXiv:2501.12540, 2025.
7. Elrefai, M.L., M.I. Khalil, and H.M. Abbas. *Multilingual Model Fine-tuning for Sentiment Analysis*. in *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. 2023. IEEE.
8. Alvi, M.B., et al., *Count me too: Sentiment analysis of roman sindhi script.* SAGE Open, 2023. **13**(3): p. 21582440231197452.
9. Brown, P.F., et al., *Class-based n-gram models of natural language.* Computational linguistics, 1992. **18**(4): p. 467–480.
10. Pisner, D.A. and D.M. Schnyer, *Support vector machine*, in *Machine learning*. 2020, Elsevier. p. 101–121.
11. Schuster, M. and K.K. Paliwal, *Bidirectional recurrent neural networks.* IEEE transactions on Signal Processing, 1997. **45**(11): p. 2673–2681.
12. Yu, Y., et al., *A review of recurrent neural networks: LSTM cells and network architectures.* Neural computation, 2019. **31**(7): p. 1235–1270.
13. Church, K.W., *Word2Vec.* Natural Language Engineering, 2017. **23**(1): p. 155–162.
14. Tifrea, A., G. Bécigneul, and O.-E. Ganea, *Poincar\'e glove: Hyperbolic word embeddings.* arXiv preprint arXiv:1810.06546, 2018.
15. Igor Santos, N.N., Luiza de Macedo Mourelle, *Sentiment analysis using convolutional neural network with fastText embeddings*, in *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. 2017, IEEE. p. 1–5.
16. Tenney, I., Das, D., Pavlick, E., *Bert rediscovers the classical nlp pipeline*. 2019, arXiv preprint arXiv:1905.05950 (2019).
17. Shaikh, S., Daudpota, S.M., Yayilgan, S.Y., Sindhu, S., *Exploring the potential of large-language models (llms) for student feedback sentiment analysis*, in *In: 2023 International conference on frontiers of information technology (FIT)*. 2023, IEEE. p. pp. 214–219.
18. Edalati, M., Imran, A.S., Kastrati, Z., Daudpota, S.M., *The potential of machine learning algorithms for sentiment classification of students' feedback on mooc*, in *Proceedings of SAI Intelligent Systems Conference*. 2021, Springer. p. pp. 11–22.
19. Kastrati, Z., A.S. Imran, and A. Kurti, *Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs.* Ieee Access, 2020. **8**: p. 106799–106810.
20. Do, H.H., Prasad, P.W., Maag, A., Alsadoon, A., *Deep learning for aspect-based sentiment analysis: a comparative review.* Expert systems with applications, 2019. **118**: p. 272–299.
21. Sindhu, I., Daudpota, S.M., Badar, K., Bakhtyar, M., Baber, J., Nurunnabi, M., *Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation.* IEEE Access, 2019. **7**: p. 108729–108741.

22. Zyout, I., and Zyout, M., *Sentiment analysis of student feedback using attention-based RNN and transformer embedding.* IAES International Journal of Artificial Intelligence (IJ-AI), 2024. **13**(2): p. 2173–2184.
23. Alvi, M.B., et al. *Hybrid classification model for twitter data-a recursive preprocessing approach*. in *2018 5th International Multi-Topic ICT Conference (IMTIC)*. 2018. IEEE.