IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 1, February, 2024, pp. 221-226
Journal homepage: http://iieta.org/journals/ria

Hybrid Approach for Automated Answer Scoring Using Semantic Analysis in Long Hindi Text

Deepender*, Tarandeep Singh Walia

School of Computer Applications, Lovely Professional University, Punjab 144411, India

Corresponding Author Email: deependerduhan6@gmail.com

https://doi.org/10.18280/ria.380122

Received: 19 September 2023
Revised: 1 December 2023
Accepted: 8 December 2023
Available online: 29 February 2024

Keywords:
automated scoring system, deep learning, LSTM, natural language processing, PSO, RNN, Roberta

ABSTRACT

In the case of the Hindi language, the technology that underpins automated scoring is still in its infancy in terms of its development. These systems have shown much better accuracy and reliability in their operations Nowadays, several studies are being carried out in addition to saving individuals money and time. with the intention of providing nuanced feedback on grammatical as well as semantic problems. This paper's main objective is to develop a hybrid methodology for automated answer scoring using semantic analysis for long Hindi text. Deep Learning and Recurrent Neural Network method have been taken into consideration throughout this research study. The ability of recurrent neural networks, to learn the temporal dependency of sequential input gives them an edge over feed forward neural networks, when it comes to the scoring of musical responses. Research work has integrated PSO and Roberta to improve accuracy. Based on the research findings, the recommended approach has been shown to outperform the currently recognized revolutionary techniques. It shows that the Hybrid PSO-Roberta based deep learning strategy performs better than the old system in terms of precision, recall, and f1 score. It reduce the amount of paperwork they need to do, teachers won't have to be concerned about any evaluation issues going away either.

1. INTRODUCTION

Using computers to evaluate content that would otherwise be assessed by humans is known as "automated scoring." Modern automated scoring systems owe a great deal to the research conducted in the area of artificial intelligence scoring. AES is the process of assigning grades to student essays written using state-of-the-art computers in a classroom setting. In addition to its use as an NLP tool, it may also be seen as a kind of educational assessment. The principle aim of the research is to develop a novel approach to simplify and automate scoring of Hindi literary works through the use of semantic analysis. Several machine learning methods are acknowledged as promising candidates for achieving this objective. It has been demonstrated that in order to increase accuracy and efficiency, filtering the dataset with an optimizer is crucial. A machine learning model must be trained using the best available data in order to guarantee optimal learning. Therefore, the suggested work has given a creative and high-performance approach to finding extended text in Hindi. The results of this research might be used to the analysis of texts containing educational statistics published in Hindi. The team's efforts are concentrated on lengthy texts written in Hindi. In order to achieve this goal, scientists are investigating several machine learning approaches. In order to get the most out of an optimizer, you need to filter the dataset. The best training data may be used if your machine learning configuration is optimized. Therefore, the suggested research has developed a very effective and perceptive method of discovering substantial Hindi literature. Research of this kind has the potential to radically transform information retrieval practices in schools and libraries that make use of Hindi-language materials.

1.1 Hindi text analysis

Hindi NLP is used in the branch of AI known as text analytics to convert the free-form text present in documents and databases into more traditional, structured data suitable for analysis or for powering ML algorithms. Examples of this include contact centre transcripts, online testimonials, surveys, and focus group notes. These raw Hindi text files contain a wealth of useful information that has yet to be extracted. These underutilized data sources may now be put to use with the help of text mining and analytics. Hindi text analysis is the process of preprocessing texts to extract information that a machine can understand. The purpose of Hindi Text Analysis is to provide structure to previously unorganized data. It's similar to attempting to make sense of mountains of paperwork by breaking it down into manageable chunks.

1.2 Automated essay scoring

Automated scoring involves using computers to evaluate tasks previously done by humans. AI techniques play a

crucial role in this process, particularly in automated essay scoring (AES) where computers assign grades to written assignments [1]. This approach is valuable in education, providing quicker feedback to students to enhance their writing skills. However, questions involving answer choices, file uploads, or diagrams still require manual grading, as there may not be a single correct solution. Subjective questions often need human graders to assess free-form responses. Automated scoring is mainly associated with using technology for tasks traditionally done by people. Automatic scoring, especially in the field of Natural Language Processing (NLP), has gained prominence [2]. It assesses students' work based on logical and semantic connections with the correct answer. Technology tends to provide more consistent ratings than human graders, whose assessments may vary based on the grader's subjectivity. Automatic short answer scoring methods are presented in Figure 1.

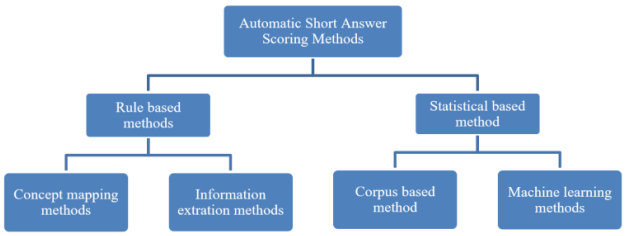


Figure 1. Methods of Automatic short answer scoring

Various studies have been offered for the purpose of data categorization and Hindi textual analysis. Some researchers have employed a supervised machine learning technique, while others have used an unsupervised one. Some studies used a multi-model approach to sentiment analysis, while others used a hybrid approach. Polynomial computing and load-balanced scheduling has also been the subject of studies. Some studies looked at using deep learning to automatically grade essays. It takes a long time for researchers to find a remedy. It has also been noted that machine learning systems' accuracy and performance might need some work. The scalability must be improved by including an optimization method. Therefore, the objective of this research is to propose a hybrid machine learning technique that combines automatic essay scoring algorithms based on natural language processing (NLP) with optimisation mechanisms to produce a more dependable, flexible, and scalable solution for Long Text Hindi categorization. The Hybrid PSO-Roberta technique has also been studied by applying a deep learning approach [3].

2. LITERATURE REVIEW

The methods used by various forms of automated scoring vary. However, the most pertinent methodological approach to this study is comprised of:

2.1 Automated essay scoring

Several case studies have been proposed for data classification, Text analysis, and computerized essay grading. There have been applications of both supervised and unsupervised machine learning. In addition, several experts have proposed a combined strategy for data classification and text analysis [4]. Ikram and Castle [5] introduced AES used a ML Strategy inspired by semantic analysis. In this study, they introduce a SA and ML-based automated essay assessment method. In order to better use feature lists, this study suggests enhancing the Coh-Metrix algorithm AES. Referential cohesiveness, lexical variety, and syntactic complexity are some of the technical criteria assessed. It also suggests four new semantic measurements, one of which was trying to figure out how much of a match there was between an essay's subject matter and its brief. The suggested AES system was put through its paces by deploying a NN-based prototype implementation to evaluate its individual and comparative performance. Results demonstrate a significant increase in neighboring accuracy from the previous study's 91% to 97.5% (with a QWK of 0.822), demonstrating a major improvement over the initial Coh-Metrix algorithm. That the additional elements and the suggested system have the potential to enhance essay graded implies that this wasa promising subject for further study. Ramesh and Sanampudi [6] focused on the development of an automatic essay grading system: a comprehensive literature study. In this study, they examine the available research on the topic of essay grading algorithms. They analyzed the limits of previous studies and emerging developments in the field of automated essay scoring via the lens of AI and ML. They found that essays were not graded according to their content's usefulness and cohesiveness.

2.2 Short answer scoring approach

In a 2012 piece, Ludwig et al. [7] spoke about how to score a short essay. Their approach breaks down ideas into their most basic components, which are model responses to closed-ended questions supplied by experts. Automatically identifying resultant ideas and their relationships allows for measurement of their interdependencies. The scoring system uses patterns to identify dependencies. For each student's response, the procedure is repeated.

Marking short essay solutions automatically depends substantially on semantic similarity, as shown by Omran and Ab Aziz [8]. Two approaches are proposed as a result of their investigation: the first is the Alternative Sentence Generator Method, which makes use of a dictionary of synonyms to generate a potential model response. The matching phase of the second hypothesis employs a hybrid algorithm consisting of the LCS, COW, and SD.

2.3 Question answering (QA) approach

The QA (Question Answering) system was a standard for automated response scoring. It involved finding specific response phrases from large document sets to answer questions. This approach allowed users to connect with computers by getting specific information as answers, rather than entire documents. Information retrieval expert Ince, E.Y., Kutlu, A. present a web based Turkish Question Answering technique in their 2021 paper [9]. Disambiguation improvements in the semantic Question Answering system were described by Hazrina et al. [10]. It's evident that ambiguity was a difficult for any SQA system. When linguistic triples are matched with numerous KB ideas, a SQA system must choose the correct interpretation via disambiguation solutions. When a linguistic triplet is not matched to a KB idea, the algorithm will suggest other words

that have similar meanings. Zupanc et al. [11] propose an expansion of previous automated essay grading systems that takes into account new semantic coherence and consistency qualities. To better estimate an essay's coherence, we developed fresh coherence qualities by mapping out its sequential components in the semantic space and analyzing the differences between them.

2.4 PSO based Roberta and LSTM approach

Shahi and Sitaula [12] reviewed the text in nepali was processed using a natural language method. In this study, they provide a comprehensive overview of NLP literature in Nepali, together with the tools that have been developed to support it. Furthermore, they use a detailed taxonomy for each NLP strategy, technique, and application task utilized in Nepali language processing. Finally, they analyze the gathered data and suggest how it might be used to further Nepali NLP studies. Our survey provides researchers with rich information on their subjects' histories and their reasons for studying NLP, paving the way for future advancements in Nepali-language NLP studies. Haseeb et al. [13] did research on the amazon customer reviews for sentiment using text mining and NLP. To save the analyst time and effort, this research uses a web-based technology to classify and analyze customer evaluations of products, saving millions of reviews from being reviewed by hand. In order to do its analysis, the system uses NLP and TA methods tailored specifically to product reviews. The text analytics programme eliminates noise and extracts emotions from the textual information. By averaging the sentimental amounts, we may arrive to the customer satisfaction score. They can look into and locate evaluations that analysts were interested in with the use of Python-based SA [14].

3. MATERIAL AND METHODS

In automated answer scoring using semantic analysis on long Hindi text, a series of essential tasks are performed to optimize the data.

These tasks encompass cleaning the text by eliminating non-alphanumeric characters and specific Hindi symbols, as well as tokenizing it, considering compound words and intricate sentence structures. Lowercasing ensures uniformity, while the removal of common Hindi stop words minimizes noise. Additionally, lemmatization or stemming captures word base forms, spelling checks enhance precision, and sentence splitting accounts for complex sentence structures. Furthermore, part-of-speech tagging assigns grammatical categories, entity recognition identifies named entities, and text normalization standardizes numerical values and dates in Hindi. Employing LSTM (Long Short-Term Memory) architecture as part of this preprocessing stage enables the model to recognize temporal dependencies and patterns within the text data, enhancing the semantic analysis's accuracy and understanding. The inclusion of Hindi word embeddings and specialized language models ensures the readiness of the text for comprehensive semantic assessment.

LSTM model: The decision to choose LSTM (Long Short-Term Memory) over other deep learning algorithms in the context of automated answer scoring and semantic analysis is based on the unique strengths of LSTMs for handling sequential data, especially when dealing with long texts [14]. LSTMs are a type of recurrent neural network (RNN) architecture that excels in capturing temporal dependencies and patterns within sequences, making them highly suitable for tasks like natural language processing, text analysis, and scoring. LSTMs can capture long-range dependencies in the data, which is crucial in understanding the context and meaning of text in longer responses [15]. This makes them effective for maintaining context and tracking semantic relationships over extended sequences. It can also handle variable-length sequences and also Preventing Vanishing Gradient problem encountered in traditional RNNs, enabling them to effectively model sequences with a more extensive context. LSTMs have a memory cell that can store and retrieve information over long periods, allowing them to capture not only the relationships between words but also the continuity and flow of ideas in a response. Different steps used for LSTM based detection Model are given in Figure 2.

ROBERTA: Similar to BERT, RoBERTa is a transformer-based language that uses self-attention to evaluate input sequences and construct phrase-level contextual representations. Different steps used for RoBERTa Model are given in Figure 3. Since RoBERTa was trained on a considerably bigger dataset than BERT, it is more effective. RoBERTa is functionally comparable to BERT but is designed differently since it employs a byte-level BPE tokenizer (like GPT-2) and a new pretraining approach. In contrast to the 16GB dataset used to train BERT, RoBERTa is trained on almost 160GB of uncompressed text.

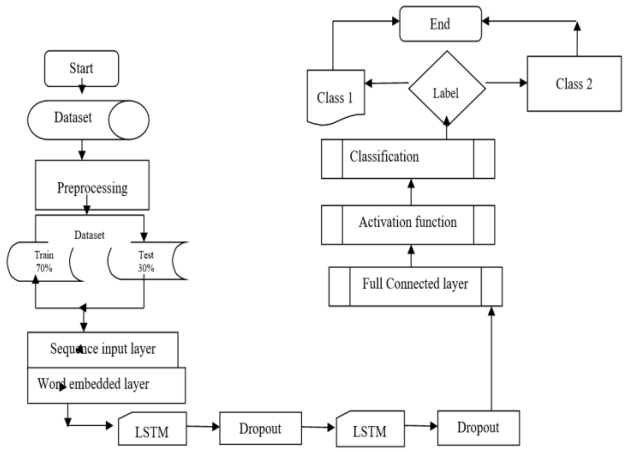


Figure 2. Flow chart for planned work

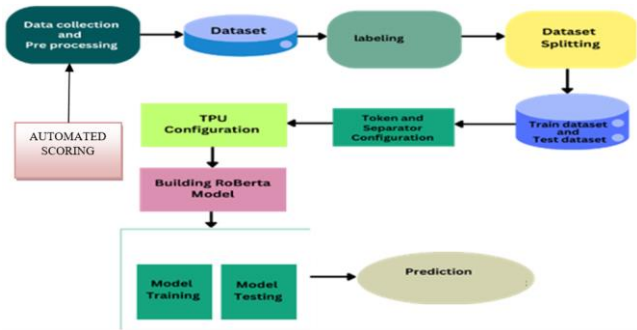


Figure 3. Roberta Model for proposed model

Flow chart for planned work is given in Figure 4. In this work, the step where the trained model's accuracy is verified is called testing. The purpose of the sample dataset is to evaluate the model's accuracy. To make predictions, a variety of datasets are processed via the network model that was

trained on the prior dataset. Model dependability is shown on the testing face. A trained network is taken into consideration and supervised during the testing phase. Next, the testing dataset is obtained and a trained network is then used to process the test in order to determine accuracy, precision, and f1-score while taking new test values into account. To begin our investigation, we will train the model on 70% of the data and test it on the remaining 30%.

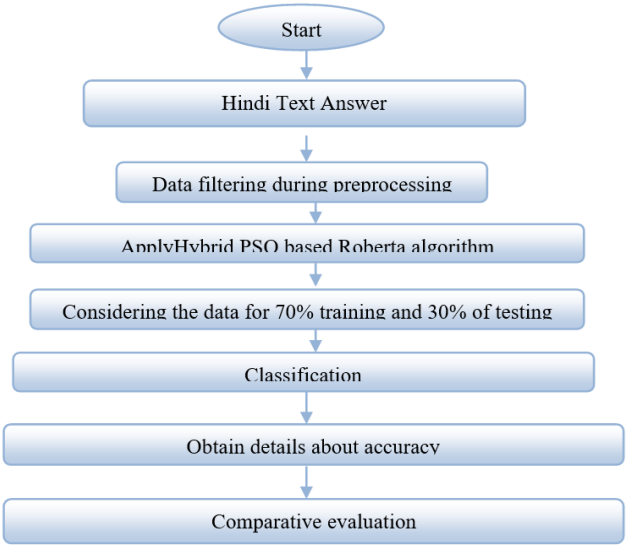


Figure 4. Flow chart for planned work

In the testing step, the accuracy of the trained model is examined so that improvements may be made if necessary. The accuracy of the model is evaluated based on the sample dataset that is collected. When performing predictions, the network model that was first trained using the prior dataset is then processed using a variety of datasets. The testing side of the model is where the dependability of it is shown. During the phase of testing known as supervision of the trained network is considered. After that, the testing dataset is taken into consideration. Following this, the data that will be tested is processed using a trained network in order to determine accuracy, f-score, and precision while taking into account fresh test values. For the sake of our study, let's train the model on the initial seventy percent of the data, and then test it on the remaining thirty percent.

Multi-model, hybrid-model, and polynomial-function support are only a few of the models that have been implemented. The issues and concerns raised by previous studies during automated scoring have been taken into account. Issues with scalability, performance, and precision were the root of the problem. The proposed model takes a creative approach to addressing problems throughout its implementation. The proposed method uses optimization, natural language processing (NLP), and classifiers in conjunction with deep learning to score essays automatically in the case of lengthy Hindi texts while minimizing performance and accuracy difficulties. After the suggested model has been constructed, its performance and accuracy will be compared. The first step in deploying a hybrid PSO-Machine learning model based on LSTM for classification is extracting answers from the dataset. After the hybrid model's parameters have been set, a training and testing set is created from the data. During the validation phase, a classification operation is carried out, and the accuracy is determined by calculating the confusion matrix. The accuracy of the suggested Hybrid method is then compared to that of the standard model.

Correlation diagram between different scoring methods: The two separate AMT systems that provide 88-note output and chrome onset output are shown on the left. Using DTW, the results are combined and synchronized to the MIDI score. We used the state-of-the-art AMT system as our starting point, rather than developing our own, since our focus was on developing a neural network-based system that generates features for automated response scoring. Nonetheless, Figure 5 shows how we made some minor adjustments to the training set.

The system's efficacy is measured by contrasting the output it produces with the outcome supplied by human raters. The results of the correlation coefficient show that the scores given by the system and the human raters are strongly associated with one another, with a value close to one indicating a positive correlation. Score agreement between a human rater and the system is quite near to the value reached between human raters. The table below displays the system-generated rankings for a sample of 50 students. The ultimate score of 50 student answers and the score provided by 5 human raters are graphically shown in Figure 6. The chart indicates that the system's output is consistent with that of human evaluators. As proof of this system's precision, a negligible outlier may be disregarded with ease. The system-generated score falls within the range of human raters' scores, demonstrating the system's accuracy and dependability.

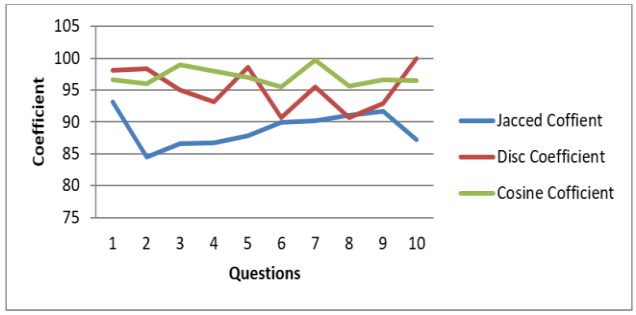


Figure 5. Correlation among various scoring techniques

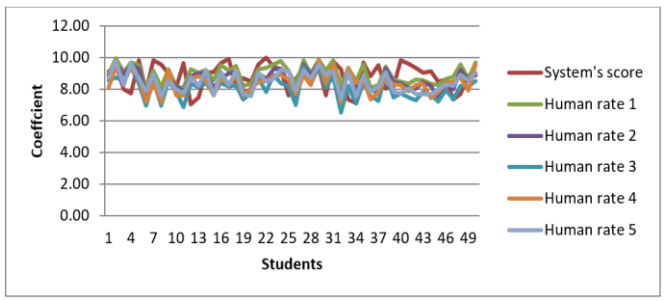


Figure 6. Comparison between 5 human raters and the system score of 50 student responses

4. RESULT AND DISCUSSION

The suggested Hybrid PSO based LSTM Model has been trained on the gathered data, which will be used to classify the answers. The resulting confusion matrices for the

unfiltered dataset are shown in Table 1, whereas those for the new method are displayed in Table 2.

The information is indicating that out of the total number of instances or cases considered, (Overall Accuracy) 88.53% were correctly classified by the model. Additionally, it provides specific information about the number of true positive (TP) cases (3541) that were correctly classified as positive.

Result: TP: 3541, Overall Accuracy: 88.53%.

Table 1. Confusion matrix for unfiltered dataset

	Choice A	Choice B	Choice C	Choice D
Choice A	336	21	24	19
Choice B	12	359	18	11
Choice C	13	11	361	15
Choice D	10	25	24	341

Table 2. Accuracy parameter for unfiltered dataset

Class	N (Truth)	N (Classified)	Accuracy	Precision	Recall	F1 Score
1	1009	1000	94.23%	0.89	0.88	0.89
2	965	1000	94.13%	0.86	0.90	0.88
3	1004	1000	94.05%	0.88	0.88	0.88
4	1022	1000	94.65%	0.90	0.88	0.89

Table 3. Confusion matrix for filtered LSTM dataset

	Choice A	Choice B	Choice C	Choice D
Choice A	368	11	12	9
Choice B	6	379	9	6
Choice C	7	6	378	9
Choice D	7	19	15	359

Result: TP: 3688, Overall Accuracy: 92.2%

Table 4. Accuracy parameter for filtered LSTM dataset

Class	N (Truth)	N (Classified)	Accuracy	Precision	Recall	F1 Score
1	1006	1000	95.95%	0.92	0.92	0.92
2	979	1000	96.08%	0.91	0.93	0.92
3	995	1000	96.08%	0.92	0.92	0.92
4	1020	1000	96.3%	0.94	0.92	0.93

The data has been utilized to train the conventional LSTM Model that will be used to categorize the replies. Results for the filtered dataset's confusion matrices can be shown in Table 3, while those using the new approach can be seen in Table 4.

The data has been utilized to train the proposed Roberta Model that will be used to categorize the replies. Results for the filtered dataset's confusion matrices can be shown in Table 5, while those using the new approach can be seen in Table 6.

Table 5. Confusion matrix for filtered dataset for Roberta

	Choice A	Choice B	Choice C	Choice D
Choice A	368	11	12	9
Choice B	6	379	9	6
Choice C	7	6	378	9
Choice D	7	19	15	359

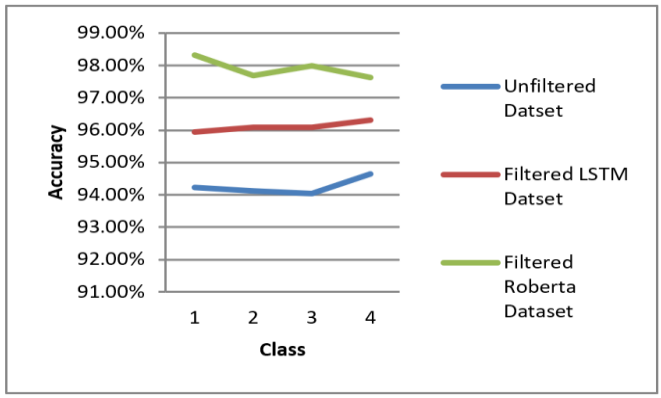
Result: TP: 1533, Overall Accuracy: 95.81%

Table 6. Accuracy parameter for filtered Roberta dataset

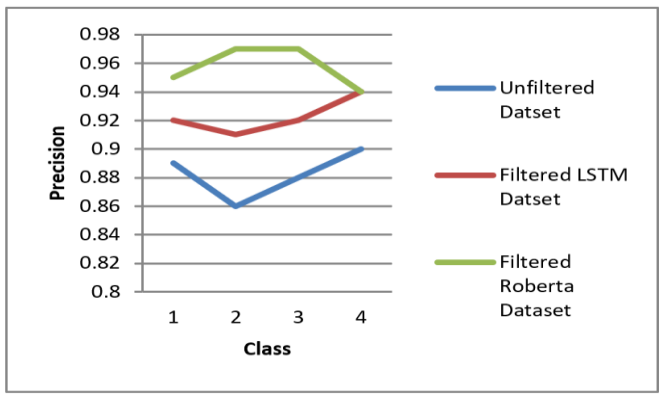
Class	N (Truth)	N (Classified)	Accuracy	Precision	Recall	F1 Score
1	388	397	98.31%	0.95	0.98	0.97
2	415	404	97.69%	0.97	0.94	0.95
3	414	408	98%	0.97	0.95	0.96
4	383	391	97.63%	0.94	0.96	0.95

4.1 Comparison analysis of accuracy parameters

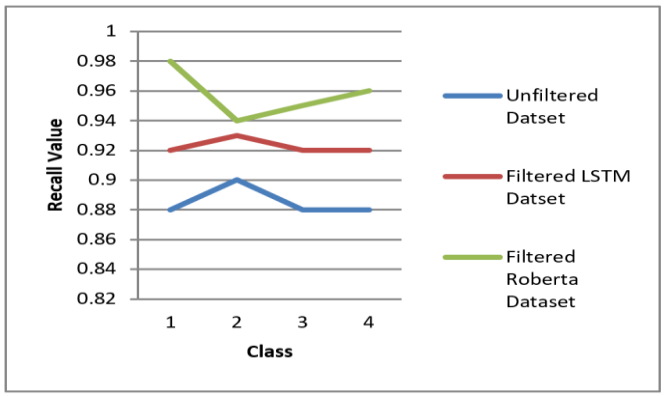
After considering the accuracy parameters from Table 2, Table 4 and Table 6, research is comparing the different parameters like precision, recall and f1 score as shown in Figure 7.



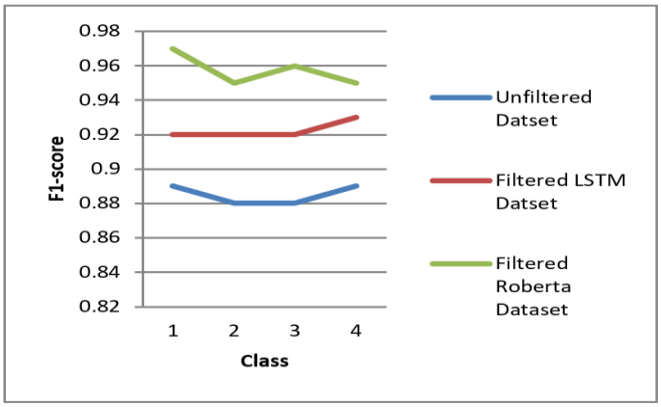
(a) Accuracy



(b) Precision



(c) Recall Value



(d) F1-Score

Figure 7. Comparison of accuracy parameter

5. CONCLUSION AND FUTURE SCOPE

Automatic answer scoring by the application of semantic analysis is offered as a new method. When using a filtered dataset, the accuracy increases from 93% to 95% suggests that the proposed approach is effective in improving the correctness of answer scoring and both the precision and recall increase (0.92 to 0.95 and 0.92 to 0.98, respectively) indicates that the model is better at correctly classifying relevant answers while reducing false positives.F1-score, which is a harmonic mean of precision and recall, also increases from 0.92 to 0.97 with an unfiltered dataset, showing an overall better balance between precision and recall. The statement that the hybrid PSO-based Roberta Model achieves an accuracy of more than 95% suggests that the proposed approach outperforms the standard method, emphasizing its effectiveness. In conclusion, the use of semantic analysis and the hybrid PSO-based Roberta Model appears to be a promising approach for automated answer scoring, as evidenced by the improvements in accuracy, precision, recall, and F1-score. The idea that automated answer scoring using hybrid approach could further enhance the work is a promising direction. It suggests that ongoing research in optimization techniques might yield even better results in the future.

REFERENCE

[1] Klebanov, B.B., Madnani, N. (2022). Automated essay scoring. Springer Nature. https://doi.org/10.1007/978-3-031-02182-4

[2] Drees, L., Kusche, J., Roscher, R. (2020). Multi-modal deep learning with sentinel-3 observations for the detection of oceanic internal waves. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2: 813-820. https://doi.org/10.5194/isprs-annals-V-2-2020-813-2020

[3] Wang, X., Wang, Y., Yuan, P., Wang, L., Cheng, D. (2021). An adaptive daily runoff forecast model using VMD-LSTM-PSO hybrid approach. Hydrological Sciences Journal, 66(9): 1488-1502. https://doi.org/10.1080/02626667.2021.1937631

[4] Kumar, V., Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. In Frontiers in Education. Frontiers Media SA, 5: 572367. https://doi.org/10.3389/feduc.2020.572367

[5] Ikram, A., Castle, B. (2020). Automated essay scoring (AES); A semantic analysis inspired machine learning approach: An automated essay scoring system using semantic analysis and machine learning is presented in this research. In Proceedings of the 12th International Conference on Education Technology and Computers, pp. 147-151. https://doi.org/10.1145/3436756.3437036

[6] Ramesh, D., Sanampudi, S.K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55(3): 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

[7] Ludwig, S., Mayer, C., Hansen, C., Eilers, K., Brandt, S. (2021). Automated essay scoring using transformer models. Psych, 3(4): 897-915. https://doi.org/10.3390/psych3040056

[8] Omran, A.M.B., Ab Aziz, M.J. (2013). Automatic essay grading system for short answers in English language. Journal of Computer Science, 9(10): 1369. https://doi.org/10.3844/jcssp.2013.1369.1382

[9] Ince, E.Y., Kutlu, A. (2021). Web-based Turkish automatic short-answer grading system. Natural Language Processing Research, 1(3-4): 46-55. https://doi.org/10.2991/nlpr.d.210212.001

[10] Hazrina, S., Sharef, N.M., Ibrahim, H., Murad, M.A.A., Noah, S.A.M. (2017). Review on the advancements of disambiguation in semantic question answering system. Information Processing & Management, 53(1): 52-69. 10.1016/j.ipm.2016.06.006

[11] Zupanc, K., Savić, M., Bosnić, Z., Ivanović, M. (2017). Evaluating coherence of essays using sentence-similarity networks. In Proceedings of the 18th International Conference on Computer Systems and Technologies, pp. 65-72. https://doi.org/10.1145/3134302.3134322

[12] Shahi, T.B., Sitaula, C. (2022). Natural language processing for Nepali text: A review. Artificial Intelligence Review, 1-29. https://doi.org/10.1007/s10462-021-10093-1

[13] Haseeb, A., Taseen, R., Sani, M., Khan, Q.G. (2023). Sentiment analysis on amazon product reviews using text analysis and natural language processing methods. In International Conference on Engineering, Natural and Social Sciences, Konya, Turkey, pp. 446-452.

[14] Nagaraj, A., Sood, M., Srinivasa, G. (2018). Real-time automated answer scoring. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Mumbai, India, pp. 231-232. https://doi.org/10.1109/ICALT.2018.00122

[15] Nasir, J.A., Khan, O.S., Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights, 1(1): 100007. https://doi.org/10.1016/j.jjimei.2020.100007