International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 8, No. 4, 2021, pp. 18-23.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Automated Essay Grading System for IELTS Using BI-LSTM

Rajan Patel [a], Neelam Yadav [b]

[a] *Department of Mechanical Engineering, TurboTech Institute of Engineering, Kolhapur, India.*
[b] *Department of Mechanical Engineering, TurboTech Institute of Engineering, Kolhapur, India.*

Abstract

The most crucial idea for text mining applications such as machine translation, text classification, text summarization, and text categorization is short-text similarity. The key question in NLP is how two words, phrases, or sentences are comparable. As a result, research on its applications is very popular. Text similarity quantifies the degree of resemblance or relationship between two words, phrases, or sentences. That similarity could be lexical or meaning-based. Semantic expert system software aims to comprehend and analyze natural language in relation to the meaning of human language and complicated, irregular, and varied intellectual environments. These days, the discipline of natural language processing (NLP) needs systematic approaches and algorithms to analyze sentence and brief text similarities. From the perspective of information theory, similarity can be defined as the similarities between two texts. The resemblance increases with ordinariness and vice versa.

Keywords: Automated Essay Grading System, Bidirectional Long Short-Term Memory, Prediction.

DOI:10.71086/IAJSE/V8I4/IAJSE0826
Received: 08 September 2021/Revised: 10 October 2021/Accepted: 05 November 2021/Published: 31 December 2021

Introduction

Natural Language Processing (NLP), another name for computational language technology, has emerged as a crucial field within artificial intelligence (AI). A huge volume of textual material has been produced in daily life. Many academics are now able to study how to analyze such data to make it clear, thorough, and effective thanks to NLP. The latest NLP applications call for an effective technique or technology to identify and correct similarities between short texts or sentences (Nguyen & Dery, 2018). language similarity, which will be applicable to web-based applications, can solve the knowledge base of a dialogue system or conversational agent by using real sentences rather than structural language patterns. Sentence similarity has been identified as an effective technique for improving the retrieval efficiency of web pages when titles are used to identify documents in the named page discovery task. The use of brief text surrounding photographs can also achieve a higher retrieval accuracy than the preference of the full content in that embedded image. Sentence similarity is used in text mining to extract private data from text databases (Mahana et al., 2012).

Applications of Natural Language Processing (NLP), such as question answering, information retrieval (IR), plagiarism detection, and extraction, heavily rely on Semantic-Text-Similarity (STS). The degree of resemblance between two texts is determined by the STS function. A technique that yields a continuous rank on a scale of 1 to 5, where 1 denotes semantic sentences that are completely independent and 5 denotes semantic equivalency. Because languages produce many synonyms and ambiguities—sentences can have complicated structures and varying lengths—STS proved a challenging task (Burstein et al., 2001). The peculiarity of a further fundamental system, like TF-IDF and Bag-Of-Words (BOW), limits its ability to play the role through word sequence and overlooks syntactic and semantic association (Attali & Burstein, 2006).

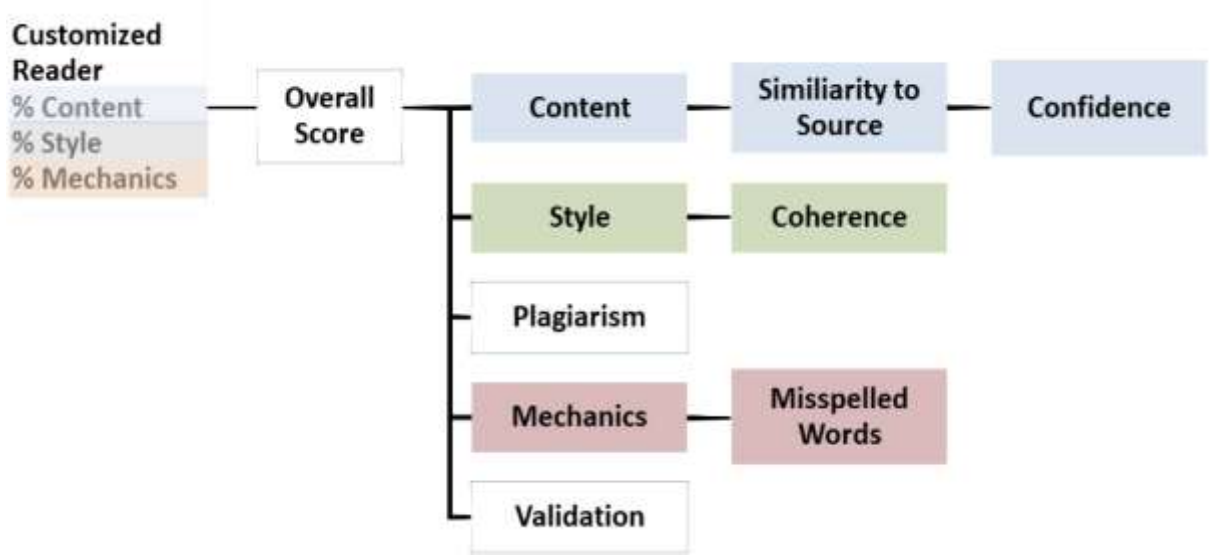


Fig. 1: Automatic Essay Scoring

Convolutional Neural Networks (CNN) use deep evaluation of words and sentences to better put both the structure and semantics in range of predicting the similarity in the phrase. Phrase Similarity has recently achieved success by using Neural Networks (NN). Thus, the paper lists numerous historic works structured using various STS systems in order to determine the advantages and disadvantages of the suggested system (Nguyen et al., 2011). Automatic Essay Scoring shown in Fig. 1.

System Model and Assumptions

Visual-semantic similarities between a sentence and an image were examined by the picture word matching test. Through the accumulation of pairwise section word text similarity, the fine-grained matching approach has recently investigated the local alignment between the sentence words and image regions, leading to advanced inference of image over text correspondence (Czaplewski, 2009). Furthermore, it was challenging to accomplish the limited configuration since a small number of important image portions can be incorrectly (or even with missing phrases) recognized. On the other hand, a different image section cannot be strictly adhering to similar words and text with a wider variety of meanings. In order to overcome these obstacles, the study (Xu et al., 2020) examined the importance

of global semantic consistency between picture areas and sentence terms as a supplemental local alignment (Slotnick, 1972). Overall Framework of the Approach shown in Fig. 2.

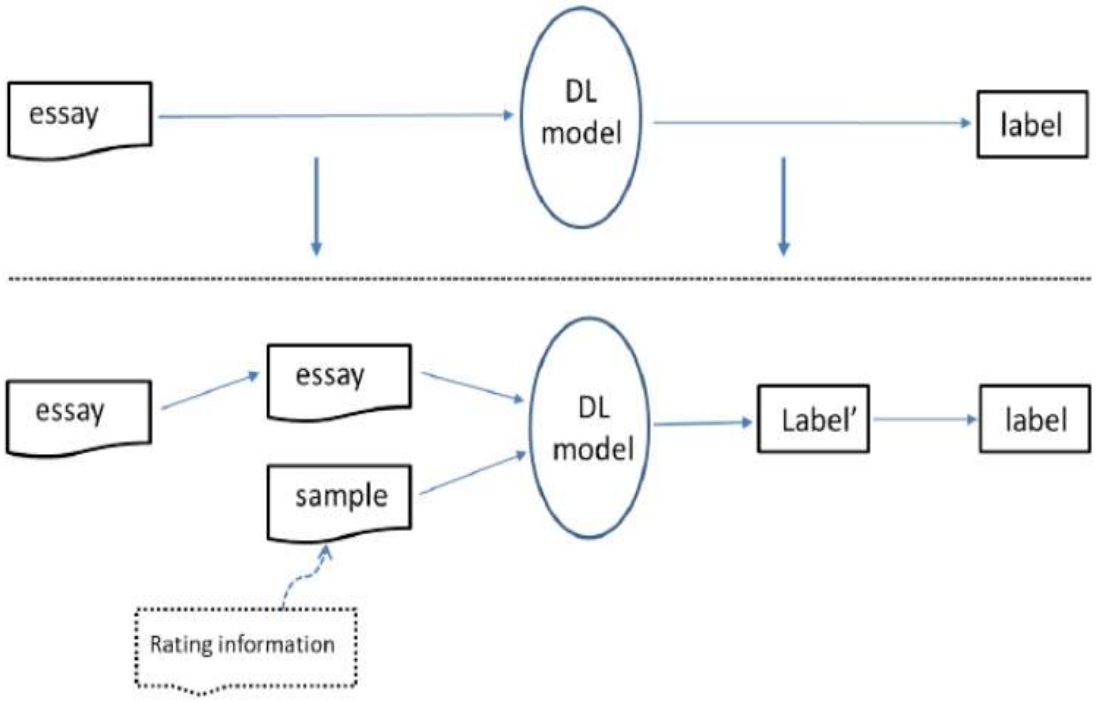


Fig. 2: Overall Framework of the Approach

Experiments

However, it's worth noting that BiLSTM entails a higher computational cost. Consequently, we undertook a comprehensive evaluation to determine the most suitable neural network for our system (Hsieh et al., 2017). Training based on LSTM Model shown in Fig. 3.

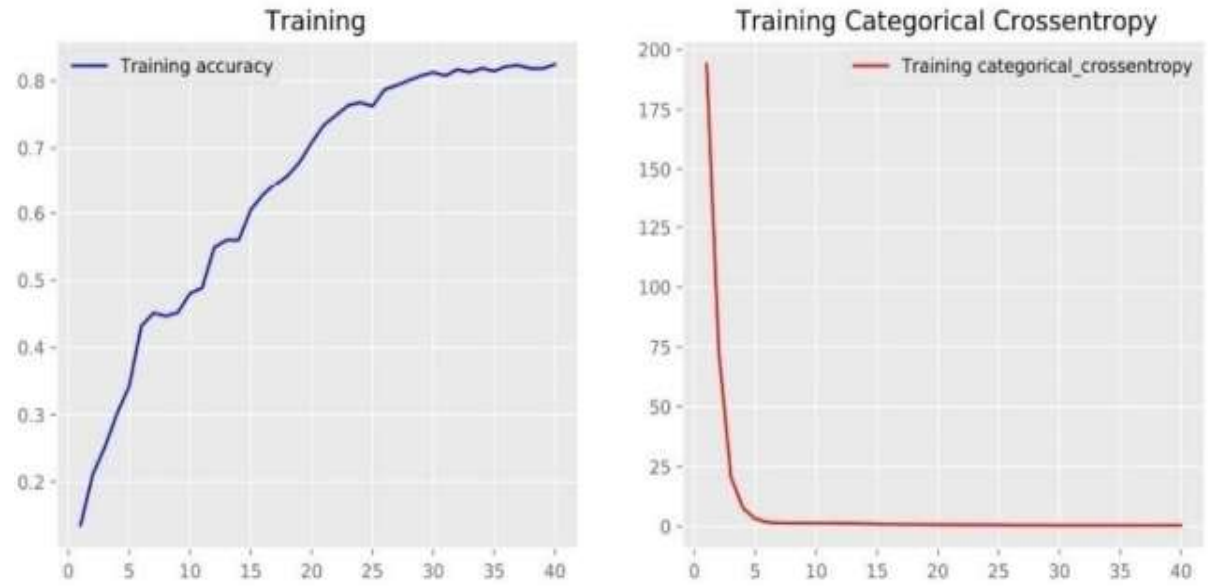


Fig. 3: Training based on LSTM Model

The average score for the LSTM was 0.69479, the average for the GRU was 0.63776, and the average for the BiLSTM was 0.67321. Furthermore, according to the top-performing models, the LSTM, GRU, and BiLSTM all had kappa values of 0.70026, 0.68525, and 0.70024, respectively. Training based on GRU Model shown in Fig. 4.

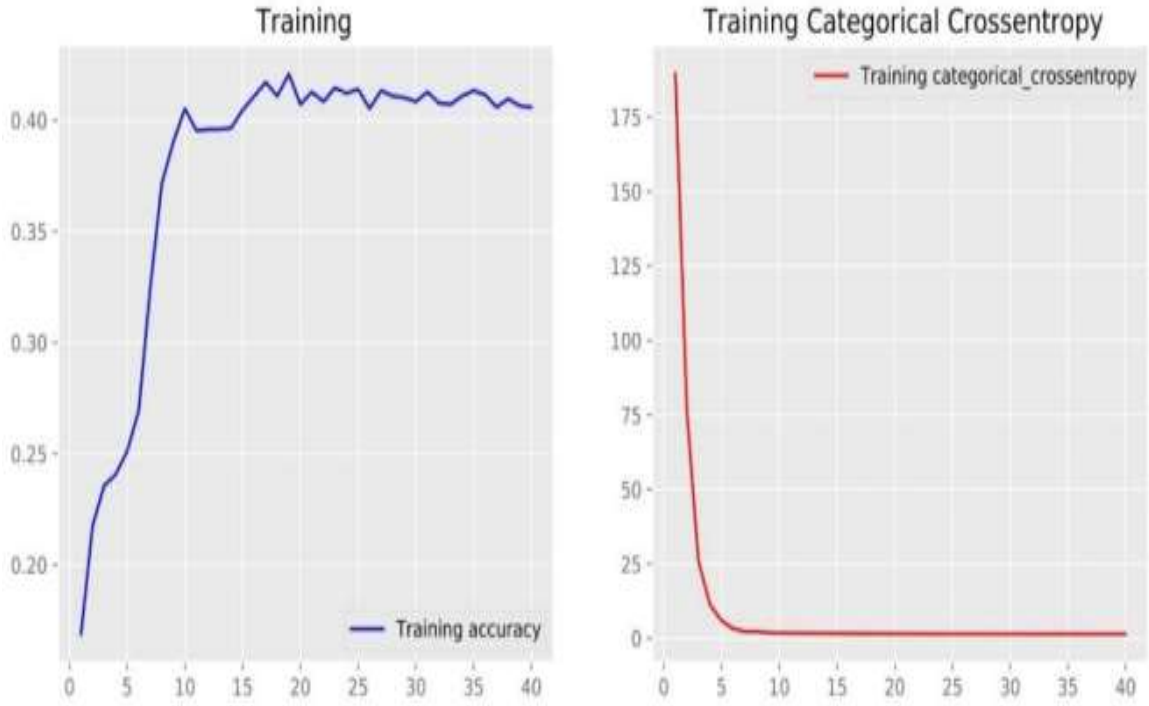


Fig. 4: Training based on GRU Model

Interestingly, the LSTM fared better overall than the other neural networks. The LSTM demonstrated the best training accuracy and the lowest training category cross-entropy for one of the eight essay sets, as shown in the accompanying figures. This pattern was also seen for the other training essay sets. Training based on Bi-Directional LSTM Model Fig. 5.

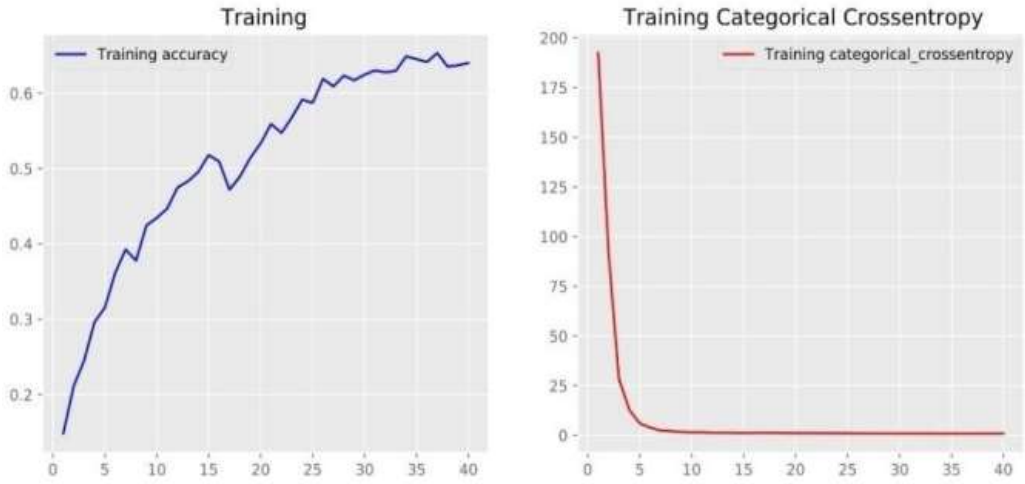


Fig. 5: Training based on Bi-Directional LSTM Model

The findings indicate that among the neural network models evaluated, LSTM emerged as the most favorable choice. In contrast, GRU exhibited poorer performance compared to the other two models. This discrepancy can be attributed to GRU's faster training time but its inadequacy for handling longer text models compared to LSTM. DR Performance Metrics shown in Fig. 6.

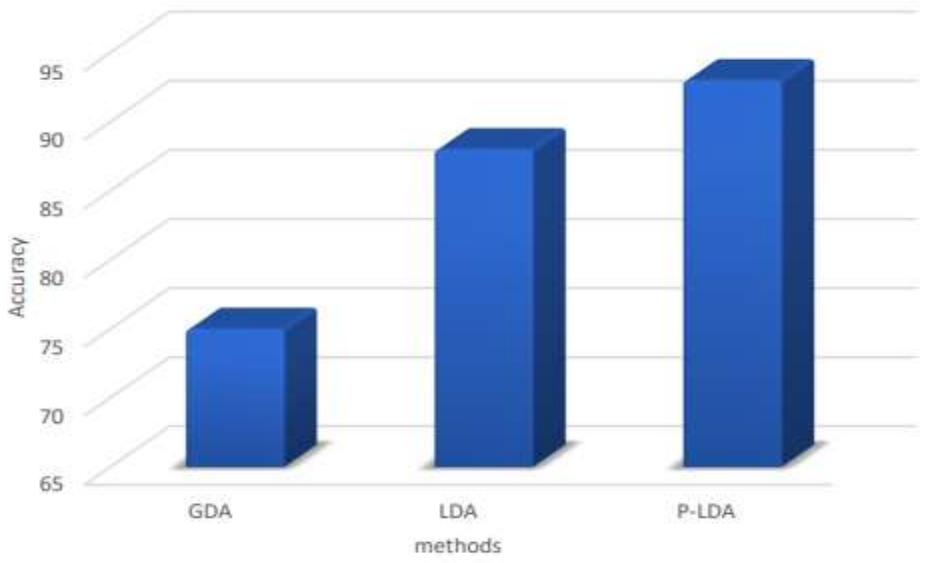


Fig. 6: DR Performance Metrics

On the other hand, BiLSTM might have preserved more information within the text, but the evaluation metric, the kappa score, assesses agreement with a human reader. Classification Performance Metrics shown in Fig. 7.

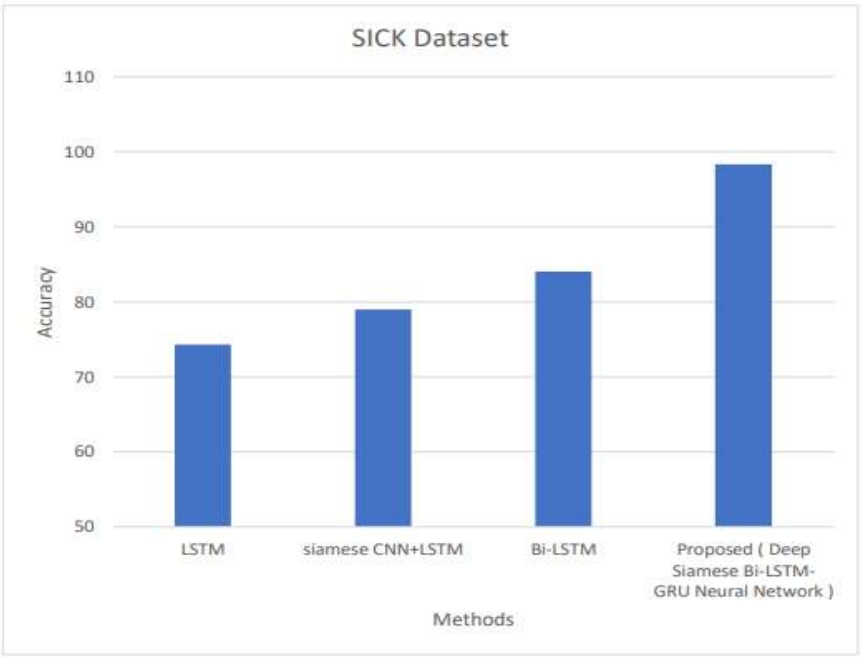


Fig. 7: Classification Performance Metrics

The subpar performance of BiLSTM may be elucidated by the fact that human graders typically read essays sequentially from beginning to end. Although BiLSTM is proficient at retaining information across the entire essay by reading it in both directions, human graders are inclined to remember information presented later in the text, mirroring the typical top-to-bottom reading pattern, which aligns more closely with LSTM's behavior.

Conclusion

The majority of earlier approaches outline research from the framework categorization to investigate how text similarity affects the representation of semantic similarity. Access to information is growing daily and quickly across

many areas. As a result, NLP applications are becoming more and more notable for making it possible to examine and get information from such data. One important component of many NLP tasks, such as text summarization and retrieval, is measuring the semantic similarity of sentences. Semantic similarity assumptions can be measured using a variety of techniques. It was noted that learning text similarity is still difficult. Thus, it becomes necessary to create a new way to get beyond the aforementioned restrictions and put it into practice for real-time applications.

References

Nguyen, H., & Dery, L. (2018). *Neural network for automated essay grading*. Stanford University.

Mahana, M., Johns, M., & Apte, A. (2012). Automated essay grading using machine learning. *Mach. Learn. Session, Stanford University*, *5*.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (2001). Enriching Automated Essay Scoring Using Discourse Marking.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Nguyen, T. D., Manh, Q. H., Minh, P. B., Thanh, L. N., & Hoang, T. M. (2011, August). Efficient and reliable camera based multiple-choice test grading system. In *The 2011 international conference on Advanced Technologies for Communications (ATC 2011)* (pp. 268-271).

Czaplewski, A. J. (2009). Computer-assisted grading rubrics: Automating the process of providing comments and student feedback. *Marketing Education Review*, *19*(1), 29-36.

Slotnick, H. B. (1972). Toward a theory of computer essay grading. *Journal of Educational Measurement*, *9*(4), 253-263.

Hsieh, K. L. C., Lo, C. M., & Hsiao, C. J. (2017). Computer-aided grading of gliomas based on local and global MRI features. *Computer methods and programs in biomedicine*, *139*, 31-38.