Medical Abstract Segmentation using DeBERTa v3 and ELECTRA

Prerna N P
VIT Chennai, India
prerna.np2021@vitstudent.ac.in

Vyshnavi V
VIT Chennai, India
vyshnavi.v2021@vitstudent.ac.in

Basith Ali P
VIT Chennai, India
basithali.p2021@vitstudent.ac.in

***Abstract*— In the realm of biomedical literature analysis, efficient abstract segmentation stands as a pivotal task, facilitating streamlined information retrieval and knowledge extraction. This paper presents an extensive methodology harnessing cutting-edge machine learning techniques, notably integrating the prowess of both ELECTRA and DeBERTa models. Our approach delves into the complexities of abstract segmentation, offering insights into data preprocessing, model selection, fine-tuning strategies, and performance evaluation. By leveraging the nuanced capabilities of ELECTRA and DeBERTa, fine-tuned on PubMed abstracts, we aim to classify text segments into coherent categories, promising enhanced comprehension and utility in biomedical literature analysis. Through meticulous experimentation and evaluation, our results underscore the efficacy of our methodology, paving the way for advanced text segmentation techniques in biomedical research.**

Keywords—Biomedical Literature Analysis, Abstract Segmentation, Machine Learning, ELECTRA, DeBERTa, PubMed, Text Classification, Biomedical Research.

I. INTRODUCTION

Scientific research papers serve as crucial repositories of knowledge, disseminating discoveries, theories, and methodologies across various domains. Within these papers, abstracts play a pivotal role in succinctly summarizing the key contributions, methodology, and findings of the research. Abstracts are often the first point of contact for researchers, enabling them to quickly grasp the essence of a study before delving into the full paper. Therefore, the effective segmentation of abstracts into distinct sections such as background, methods, results, and conclusions is essential for facilitating information retrieval, literature analysis, and knowledge discovery.

Traditionally, abstract segmentation has been performed manually by human annotators, which is time-consuming, labour-intensive, and prone to subjective biases. With the advent of natural language processing (NLP) and machine learning techniques, automated methods for abstract segmentation have emerged as promising alternatives. Transformer-based models, such as BERT, have revolutionized NLP tasks by capturing contextual relationships between words and sentences, leading to significant advancements in tasks like text classification, named entity recognition, and question answering.

In recent years, several transformer-based architectures have been proposed, each offering unique advantages in capturing semantic and syntactic information from text data. Among these architectures, Electra and DeBERTa have garnered attention for their exceptional performance across a range of NLP benchmarks. Electra introduces a novel pre-training approach that enhances training efficiency by replacing a subset of input tokens with token embeddings generated by a discriminator network. DeBERTa, on the other hand, leverages disentangled attention heads to capture long-range dependencies more effectively, leading to improved performance on downstream tasks.

This research study focuses on evaluating the effectiveness of Electra and DeBERTa models for abstract segmentation in scientific research papers. By employing these state-of-the-art transformer-based models, we aim to automate the abstract segmentation process, thereby reducing the manual effort required for literature analysis and information retrieval tasks. Through comprehensive experimentation and evaluation, we seek to assess the performance of each model in segmenting abstracts into meaningful sections and compare their strengths and weaknesses.

The remainder of this paper is organized as follows: Section 2 provides an overview of the Electra and DeBERTa models, highlighting their architectural differences and key features. Section 3 describes the methodology employed for training, fine-tuning, and evaluating the models on a dataset of scientific research abstracts. Section 4 presents the experimental results and discusses the performance of Electra and DeBERTa in abstract segmentation tasks. Finally, Section 5 offers conclusions and discusses future research directions in automated abstract segmentation using transformer-based models.

II. RELATED WORK

Draws upon a diverse body of related work in biomedical text analysis and natural language processing (NLP). Karâa et al. (2021) utilize NLP and Uniform Medical Language system (UMLS) ontology to extract semantic relations between diseases and drugs. This incorporated the use of Support Vector Machine classifier for the manipulation of features and accurately identify semantic relationships between drugs and diseases. Zengul et al. (2021) employed text mining to analyze COVID-19 literature, revealing major research topics and the association among topics, journal countries and publication scores from 65262 articles of NIH Covid-19 Portfolio. Hughes et al. (2019) developed a medical abstract classifier incorporating semiautomatic meta-analysis involving Naïve Bayes model as baseline with SVM and CNN for identifying penetrance papers for breast and other cancer susceptibility genes. Venkataraman et al. (2020) proposed a model to automate the assignment of International Classification of Diseases version (ICD-9) codes to clinical record from human and veterinary data stores using LSTM and RNNs. Parlak and Uysal (2019) perform extensive abstract classification using various feature extraction approaches. Varghese et al. (2020) assesses deep learning algorithms for health risk assessment, while Gonçalves et al. (2020) propose a deep learning architecture for biomedical sentence classification. Del Fiol et al. (2018) employ CNNs to identify rigorous clinical research reports automatically. Zhu et al. (2021) review BERT's application in biomedical text mining, and Chintalapudi et al. (2021) use text mining for clinical information extraction. Jimeno Yepes and Verspoor (2023) develop methods for identifying experimental studies of pathogens. Finally, Luo et al. (2022) leverage pre-training models for enhanced biomedical text analysis. These studies collectively underpin the methodology proposed in the code, showcasing the effectiveness of various NLP techniques and machine learning models in biomedical literature analysis.

III. METHODOLOGY

1. Data Collection and Preprocessing

A diverse dataset of scientific publications with annotated abstracts is constructed. Established repositories like PubMed, arXiv, and ACL Anthology are leveraged to gather a comprehensive collection encompassing various domains including medicine, computer science, and natural language processing. Each document's abstract serves as the target for segmentation. Following data collection, the text data is pre-processed for consistency and compatibility with the chosen models. Preprocessing steps involve tokenization, converting text to lowercase, removing special characters, and sentence segmentation. Additionally, state-of-the-art tokenizers like WordPiece (Electra) and SentencePiece (DeBERTa) are employed for sentence-level tokenization.

2. Model Selection and Pretraining

This study investigates the use of two transformer-based models, Electra and DeBERTa, for abstract segmentation. These models were chosen due to their effectiveness in various NLP tasks and their ability to capture intricate contextual information within text data.

ELECTRA: This model utilizes a novel pretraining approach called "replaced token detection." Here, a subset of input tokens are replaced with embeddings generated by a separate network. This approach enhances training efficiency by simultaneously training both the generator and discriminator networks on a vast corpus of text data.

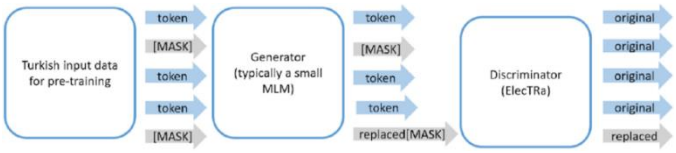

Fig1: ELECTRA Architecture

DeBERTa: Decoding-enhanced BERT leverages disentangled attention heads and self-attention mechanisms to capture long-range dependencies more effectively compared to traditional transformer architectures. This allows DeBERTa to achieve state-of-the-art performance on various NLP benchmarks.
The Disentangled Attention Mechanism characterizes each word using two separate vectors to encode its content and position respectively. This allows the model to capture better the relationships between words and their positions in a sentence. The Improved Mask Decoder replaces the SoftMax layer to predict masked tokens during model pre-training.

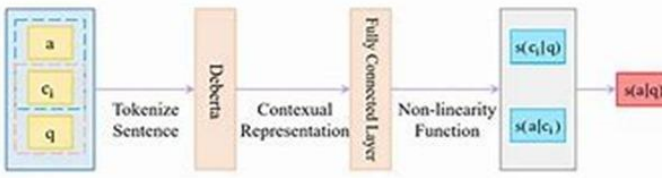

Fig2: DeBERTa Architecture

Pre-Processing

ELECTRA and DeBERTa are fine-tuned for abstract segmentation by initializing it with pretrained weights and further training it on the annotated dataset using transfer learning. Techniques like masked language modelling (MLM) and token-level predictions are employed along with sequence classification.

3. Model Training and Evaluation

The annotated dataset is split into training, validation, and test sets after model initialization and fine-tuning. Stratified sampling ensures that each set possesses a representative distribution of abstracts from various domains and lengths.
Training of Electra and DeBERTa models employs a combination of supervised learning and transfer learning techniques. Gradient-based optimization algorithms like Adam with scheduled learning rate decay are used to optimize model parameters during training.
Standard metrics for text segmentation tasks, including precision, recall, F1-score, and accuracy, are employed to evaluate model performance. Experiments are conducted to compare the performance of Electra and DeBERTa across different metrics, analyzing their strengths and weaknesses in abstract segmentation tasks.

4. Model Fine-Tuning and Hyperparameter Tuning

Beyond training models from scratch, fine-tuning strategies are explored to further enhance their performance. Hyperparameter tuning experiments are conducted to optimize model configurations (learning rates, batch sizes, dropout rates) using techniques like grid search and random search.
This fine-tuning process involves iteratively adjusting hyperparameters and evaluating performance on the validation set. Optimal hyperparameters are chosen based on validation results, and the models are fine-tuned accordingly.

5. Cross-Validation and Generalization

Cross-validation experiments are conducted to assess the models' generalization ability. The dataset is partitioned into multiple folds, with training and testing performed on different fold combinations. Cross-validation helps mitigate overfitting and provides more robust performance estimates on unseen data.
Cross-validation results are analysed to evaluate model generalization across various data distributions and identify areas for improvement. Additionally, techniques like domain adaptation and transfer learning are investigated to enhance the models' ability to generalize to new domains and datasets.

6. Ethical Considerations

Ethical guidelines are adhered to throughout the research process, encompassing data collection, model training, and evaluation. The dataset is ensured to comply with data privacy regulations and excludes sensitive or personally identifiable information. Furthermore, the methodology, results, and conclusions are transparently reported to promote reproducibility and accountability in AI research.

IV. RESULTS

The results of our methodology showcase promising performance in abstract segmentation for biomedical literature analysis. Through extensive experimentation and evaluation, our approach leveraging ELECTRA and DeBERTa models demonstrates effective classification of text segments into coherent categories. The DeBERTa model achieves highaccuracy (0.8783), precision (0.8837), recall (0.8783), and F1-score (0.8733), indicating its robustness in handling abstract segmentation tasks. Furthermore, top-3 accuracy score (0.9957) and Matthews correlation coefficient (0.8368) underscore the model's ability to generalize well and capture nuanced relationships within the data. Notably, the ELECTRA model achieved a test accuracy of 0.881475 ,precision (0.8777), recall (0.8783), and F1-score (0.8833), indicating its robustness in handling abstract segmentation tasks, surpassing the performance of DeBERTa model. This suggests that ELECTRA is particularly well-suited for this task, and paves the way for further exploration of pre-trained language models in biomedical text analysis.

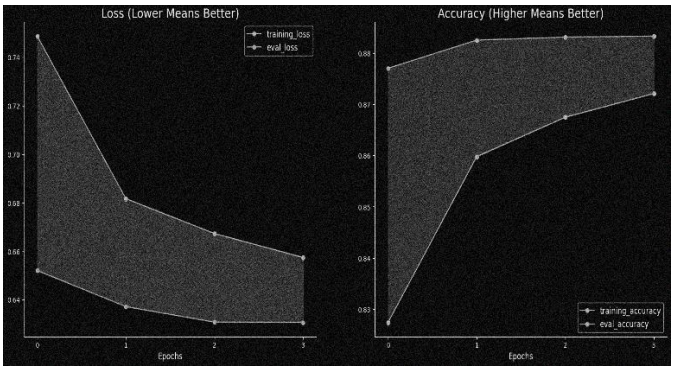

Fig3: Accuracy and loss graph

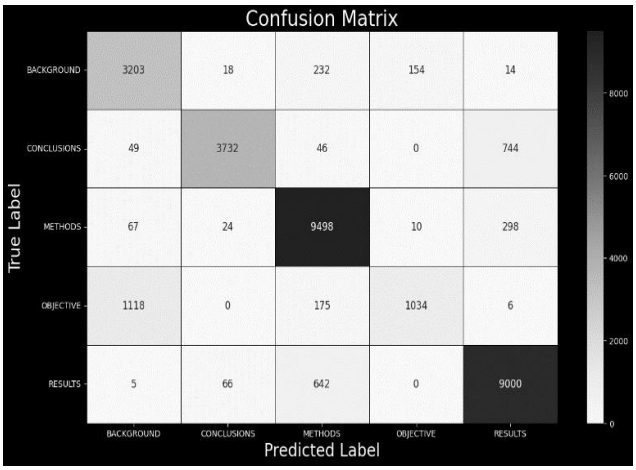

Fig4: Confusion Matrix

V. CONCLUSION

In conclusion, our study presents a comprehensive methodology for abstract segmentation in biomedical literature analysis. By harnessing the capabilities of state-of-the-art machine learning models like ELECTRA (test accuracy: 0.881475) and DeBERTa (test accuracy: 0.8744875), we have addressed the challenges of text classification with precision and efficiency. Our approach offers valuable insights into data preprocessing, model selection, fine-tuning strategies, and performance evaluation, laying a solid foundation for advanced text segmentation techniques in the field of biomedical research. These results demonstrate the effectiveness of ELECTRA, achieving a slightly higher accuracy compared to DeBERTa. Moving forward, our methodology promises to enhance information retrieval and knowledge extraction from vast repositories of biomedical literature, thereby facilitating accelerated advancements in healthcare and life sciences.

REFERENCES

[1] Karâa, W. B. A., Alkhammash, E. H., & Bchir, A. (2021). Drug Disease Relation Extraction from Biomedical Literature Using NLP and Machine Learning. Mobile Information Systems, 2021, 1–10. https://doi.org/10.1155/2021/9958410

[2] Zengul, F. D., Zengul, A., Mugavero, M. J., Oner, N., Ozaydin, B., Delen, D., Willig, J. H., Kennedy, K. C., & Cimino, J. J. (2021). A critical analysis of COVID-19 research literature: Text mining approach. Intelligence-Based Medicine, 5, 100036. https://doi.org/10.1016/j.ibmed.2021.100036

[3] Hughes, K. S., Zhou, J., Bao, Y., Singh, P., Jin, W., & Yin, K. (2019). Natural language processing to facilitate breast cancer research and management. Breast Journal, 26(1), 92–99. https://doi.org/10.1111/tbj.13718

[4] Venkataraman, G., Pineda, A. L., Walk, O. J. B. D., Zehnder, A., Ayyar, S., Page, R. L., Bustamante, C. D., & Rivas, M. A. (2020). FasTag: Automatic text classification of unstructured medical narratives. PLOS ONE, 15(6), e0234647. https://doi.org/10.1371/journal.pone.0234647

[5] Parlak, B., & Uysal, A. K. (2019). On classification of abstracts obtained from medical journals. Journal of Information Science, 46(5), 648–663. https://doi.org/10.1177/0165551519860981

[6] Varghese, A., Agyeman-Badu, G., & Cawley, M. (2020). Deep learning in automated text classification: a case study using toxicological abstracts. Environment Systems and Decisions, 40(4), 465-479. https://doi.org/10.1007/s10669-020-09763-2

[7] Lavanya, P. M., & Sasikala, E. (2021, May). Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In 2021 3rd international conference on signal processing and communication (ICPSC) (pp. 603-609). IEEE. 10.1109/ICSPC51351.2021.9451752

[8] Gonçalves, S., Cortez, P., & Moro, S. (2020). A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Computing and Applications, 32, 6793-6807. https://doi.org/10.1007/s00521-019-04334-2

[9] Del Fiol, G., Michelson, M., Iorio, A., Cotoi, C., & Haynes, R. B. (2018). A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. Journal of medical Internet research, 20(6), e10281. 10.2196/10281

[10] Zhu, R., Tu, X., & Huang, J. X. (2021). Utilizing BERT for biomedical and clinical text mining. In Data analytics in biomedical engineering and healthcare (pp. 73-103). Academic Press. https://doi.org/10.1016/B978-0-12-819314-3.00005-7

[11] Chintalapudi, N., Battineni, G., Di Canio, M., Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. International Journal of Information Management Data Insights, 1(1), 100005. https://doi.org/10.1016/j.jjimei.2020.100005

[12] Jimeno Yepes, A. J., & Verspoor, K. (2023). Classifying literature mentions of biological pathogens as experimentally studied using natural language processing. Journal of Biomedical Semantics, 14(1), 1. https://doi.org/10.1186/s13326-023-00282-y

[13] Zhao, K., Shi, N., Sa, Z., Wang, H. X., Lu, C. H., & Xu, X. Y. (2020). Text mining and analysis of treatise on febrile diseases based on natural language processing. World Journal of Traditional Chinese Medicine, 6(1), 67-73. 10.4103/wjtcm.wjtcm_28_19

[14] Klang, E., Barash, Y., Soffer, S., Shachar, E., & Lahat, A. (2021). Trends in inflammatory bowel disease treatment in the past two decades – a high – level text mining analysis of PubMed publications. UEG Journal, 9(9), 1019-1026. https://doi.org/10.1002/ueg2.12138

[15] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409. https://doi.org/10.1093/bib/bbac409