Mediterranean Smart Cities Conference (MSCC 2024)
2-4 May, 2024, Martil, Morocco

DeBERTa-Enhanced Extreme Multi-Label Classification for Biomedical Articles

1st Oussama Ndama
DSAI2S Research Team, C3S Laboratory
FST of Tangier, Abdelmalek Essaâdi University
Tetouan, Morocco
oussama.ndama@etu.uae.ac.ma
0009-0005-1303-7114

2nd Ismail Bensassi
DSAI2S Research Team, C3S Laboratory
FST of Tangier, Abdelmalek Essaâdi University
Tetouan, Morocco
bensassi.ismail@gmail.com
0009-0004-6712-2456

3rd El Motkhtar En-Naimi
DSAI2S Research Team, C3S Laboratory
FST of Tangier, Abdelmalek Essaâdi University
Tetouan, Morocco
en-naimi@uae.ac.ma
0000-0002-5846-5252

Abstract— **This study introduces a novel approach to the extreme multi-label classification of biomedical articles, leveraging the advanced capabilities of the DeBERTa (Decoding-enhanced BERT with Disentangled Attention) model. Aimed at addressing the challenges of classifying research articles into Medical Subject Headings (MeSH) labels, our methodology encompasses a comprehensive preprocessing strategy, model adaptation for multi-label classification, and a rigorous evaluation framework. Our DeBERTa-enhanced architecture demonstrates notable success in achieving a Test F1 Accuracy of 85.58%, underscoring its effectiveness in balancing precision and recall across a diverse array of MeSH categories. Through a detailed analysis of model performance across various labels, we identify strengths and highlight areas for future improvement. This study not only contributes to the field of natural language processing (NLP) applications in biomedical literature but also sets the stage for further advancements in multi-label classification techniques.**

Keywords—DeBERTa; Multi-Label Classification; Text Classification; Transformer Models; Natural Language Processing

I. INTRODUCTION

The classification of biomedical literature into predefined categories is a critical task in the field of bioinformatics, facilitating efficient information retrieval, literature review, and research discovery. With the exponential growth of biomedical research articles [1], automated classification systems have become indispensable. Among the challenges faced in this domain, extreme multi-label classification, where each document may belong to multiple categories simultaneously, is particularly daunting due to the vast and imbalanced label space represented by MeSH labels [2].

Recent advancements in NLP, particularly the development of transformer models like BERT (Bidirectional Encoder Representations from Transformers) and its variants, have shown promising results in text classification tasks [3]. DeBERTa, with its decoding-enhanced attention mechanism, offers significant improvements over traditional BERT models in understanding complex language structures and semantics [4]. This study explores the application of DeBERTa in the extreme multi-label classification of biomedical articles, aiming to harness its sophisticated language comprehension capabilities to improve classification accuracy and efficiency.

Our research is structured into a comprehensive methodology that includes the detailed description of the dataset used, an overview of multi-label classification challenges and strategies, a deep dive into the DeBERTa model, and the presentation of our proposed approach tailored for biomedical article classification. By evaluating our model across various performance metrics, we provide insights into its efficacy and delineate a path for future research endeavors aimed at refining and expanding the capabilities of NLP models in the biomedical domain. Through this study, we contribute to the growing body of knowledge in biomedical text mining and offer a framework for future advancements in automated literature classification.

II. LITERATURE REVIEW

In the rapidly evolving field of biomedical text classification, recent studies have underscored the potential of transformer-based models to address the challenges of extreme multi-label classification and automated coding tasks. Liu et al. [5] explored the effectiveness of pretrained Transformer models for International Classification of Diseases (ICD) coding, introducing a novel model, XR-LAT, to tackle the complexities of extreme label sets and long text classification. Their findings highlight the advanced capabilities of Transformer models in achieving state-of-the-art performance on benchmark datasets, setting new precedents for automated ICD coding.

Similarly, Xiong et al. [6] proposed a novel two-stage framework, XRR, for extreme multi-label text classification, addressing computational efficiency and information loss issues present in existing methods. Their

approach combines candidate retrieval strategies with a deep ranking model, demonstrating superior performance across multiple datasets and advancing the field of XMTC.

Chen et al. [7] introduced LitMC-BERT, a transformer-based multi-label classification method tailored for the curation of COVID-19 literature. By capturing label-specific features and correlations between label pairs, LitMC-BERT achieves remarkable improvements in micro-F1 and instance-based F1 scores, underscoring the transformative impact of NLP technologies in managing the deluge of pandemic-related literature.

Ibrahim et al. [8] presented GHS-NET, a generic hybridized shallow neural network that leverages both convolutional and recurrent neural networks for multi-label biomedical text classification. Their methodology showcases significant performance enhancements in classifying diverse genres of biomedical text, from literature to clinical notes, highlighting the versatility and robustness of their approach.

Lastly, Abdillah et al. [9] explored ensemble-based methods for multi-label classification on biomedical question-answer data, illustrating the efficacy of combining heterogeneous deep learning and machine learning models. Their ensemble approach outperforms single models, offering a promising direction for enhancing accuracy in biomedical data classification.

Together, these studies form a rich tapestry of research that not only demonstrates the advanced capabilities of transformer-based and ensemble methods in biomedical text classification but also sets the stage for future innovations in this vital area of bioinformatics.

III. METHODOLOGY AND MATERIALS

In the "Methodology and Materials" section of our study, we meticulously outline the foundational elements that underpin our research. This section is structured into four critical subsections: Data description, Multi-label classification, DeBERTa, and The Proposed approach. Each subsection is dedicated to providing a detailed overview of the respective aspect of our methodology—from the intricacies of the dataset utilized to the theoretical and practical applications of DeBERTa in our study. This comprehensive delineation of methods and materials ensures a transparent and replicable research process, laying the groundwork for our exploration into advanced NLP techniques for biomedical literature classification.

A. *Data description*

The dataset employed in our study, is a meticulously curated collection of approximately 50,000 research articles from the PubMed repository, each annotated with MeSH labels by biomedical experts [10]. These articles are each characterized by an extensive array of 10 to 15 MeSH major labels, reflecting the specialized topics covered in biomedical research [11]. The dataset presents a unique challenge due to its extremely large output space and severe label sparsity issues, stemming from the vast number of MeSH major labels. To address these challenges, the dataset has been processed to simplify the classification task by mapping each label to its root category, thus allowing for a hierarchical classification approach. This hierarchical structure is crucial for our study as it organizes the MeSH labels into 14 broad root categories, namely: ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'L', 'M', 'N', 'Z'], facilitating a structured approach to understanding and classifying biomedical themes.

This innovative approach not only simplifies the initial classification task by focusing on these broader categories but also preserves the relational dynamics among the labels, addressing the inherent complexities associated with the extensive number of MeSH majors assigned to each document and enhancing the learning algorithms' ability to accurately predict relevant labels. This methodology section underpins our study's approach to tackling the nuanced thematic landscape of biomedical research articles through a DeBERTa-enhanced extreme multi-label classification framework.

Figure 1 illustrates the distribution of articles across various root labels, highlighting the prevalence of specific topics within the biomedical article dataset.

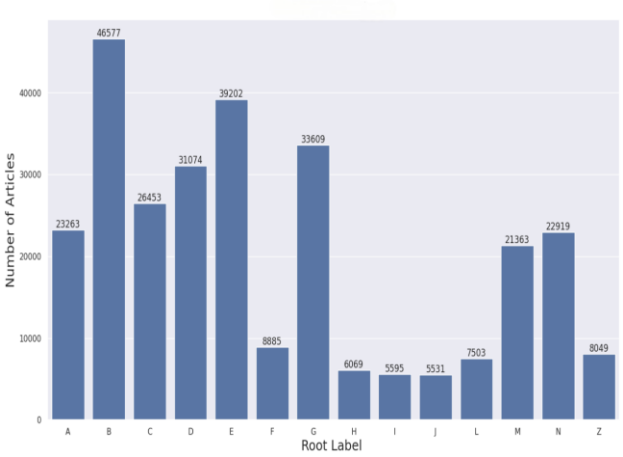

Figure 1. Number of articles per root label

B. *Multi-label classification*

Multi-label classification is a pivotal machine learning technique where an instance can be simultaneously associated with multiple labels from a predefined set. Unlike traditional single-label classification tasks, where each instance is assigned to a single category, multi-label classification acknowledges the complexity and richness of real-world data by allowing for the assignment of multiple relevant categories to each instance [12,13]. This approach is particularly relevant in domains where the subjects under consideration exhibit multiple attributes or belong to several categories simultaneously, such as in the annotation of biomedical articles with MeSH labels. Each article might cover a range of topics, necessitating the assignment of several MeSH labels to accurately capture its content. The challenge in multi-label classification lies not only in predicting the correct set of labels for each instance but also in managing the interdependencies and correlations among the labels [14]. Advanced algorithms and models, such as DeBERTa, are often employed to tackle this complexity, leveraging deep learning and natural language processing techniques to enhance

prediction accuracy and handle the high-dimensional space of labels efficiently. Multi-label classification thus plays a crucial role in organizing, retrieving, and understanding complex datasets, enabling more nuanced analyses and insights in fields like biomedical research [15].

C. DeBERTa: Decoding-enhanced BERT with Disentangled Attention

DeBERTa, which stands for "Decoding-enhanced BERT with Disentangled Attention," represents a significant advancement in the field of natural NLP and machine learning [16]. This model introduces a novel architecture that improves upon the BERT framework by incorporating two key innovations: disentangled attention mechanism and enhanced mask decoder. The disentangled attention mechanism separates the representation of content and position within the model, allowing DeBERTa to more accurately understand the importance of word positions and the contextual relationships between words in a sentence [17]. This feature enables the model to capture the nuances of language, including syntax and semantic meanings, more effectively than its predecessors.

Additionally, DeBERTa enhances the pre-training phase with a mask decoder that predicts not only the masked tokens but also their positions, further refining the model's understanding of language structure and context. These innovations allow DeBERTa to achieve superior performance across a wide range of NLP tasks, including natural language understanding, natural language inference, and question-answering challenges [18].

In the context of our study, DeBERTa's sophisticated understanding of language nuances and structure makes it exceptionally suited for dealing with the complexities of biomedical literature. Its ability to discern and encode the subtle differences in biomedical terminologies, their relationships, and contexts allows for more accurate classification of articles into multiple, relevant MeSH labels. This makes DeBERTa an ideal choice for enhancing extreme multi-label classification tasks, where the goal is to accurately assign a large set of labels to each document, navigating through the intricacies of biomedical knowledge.

D. The proposed approach

In our study, we present a sophisticated architecture that adeptly addresses the challenges associated with the multi-label classification of biomedical literature. The foundation of our approach is the utilization of the DeBERTa model, renowned for its exceptional capability in understanding complex language representations. Our methodology is meticulously designed to harness the full potential of DeBERTa, tailored specifically for the nuanced demands of biomedical text analysis.

The preprocessing stage is critical to our approach, where biomedical articles are encoded using the `DebertaTokenizer`. This tokenizer is adept at handling the specialized vocabulary of biomedical texts, ensuring that the model can accurately interpret the context and semantics of the articles [19]. To accommodate the extensive scope of the dataset, which comprises approximately 50,000 PubMed articles, we enforce a `max_length` parameter of 128 during tokenization. This length is strategically chosen to balance the need for sufficient contextual information against computational constraints.

Our model configuration leverages the `microsoft/deberta-base` variant of DeBERTa, a decision motivated by the variant's proven effectiveness in balancing computational efficiency with advanced language comprehension capabilities [20]. The model is fine-tuned for the task of multi-label classification, with adjustments made to align with the unique structure of MeSH labels present in the dataset.

The architecture employs a binary cross-entropy loss function, `BCEWithLogitsLoss`, which is specifically chosen for its suitability in multi-label classification scenarios [21]. This loss function, coupled with the `AdamW` optimizer set at a learning rate of `6e-6`, forms the backbone of our training regimen [22]. Such a configuration is pivotal in optimizing the model's parameters, ensuring it can navigate the intricacies of label sparsity and imbalance prevalent in biomedical literature classification tasks.

Training is conducted using a batch size of 32, optimized through `DataLoader` for efficient processing and resource utilization. This setup ensures that our model not only learns effectively from the data but does so in a manner that is computationally sustainable [23]. We underscore the importance of hardware acceleration, as our code is designed to detect and utilize GPU capabilities, thereby significantly enhancing the training and inference processes.

Evaluation of the model's performance is conducted through the lenses of F1 score and accuracy, metrics that are indispensable in the realm of multi-label classification. These metrics offer a comprehensive view of the model's efficacy, providing insights into its ability to accurately classify articles into multiple MeSH labels.

In summary, our study introduces a robust and innovative application of the DeBERTa model, tailored for the extreme multi-label classification of biomedical articles. Through strategic preprocessing, model configuration, and rigorous evaluation, we demonstrate the model's prowess in navigating the complex landscape of biomedical research literature, setting a precedent for future endeavors in the application of deep learning models to the domain of biomedical text analysis.

IV. RESULTS AND DISCUSSION

In this section, we present a detailed evaluation of our DeBERTa-enhanced architecture for extreme multi-label classification of biomedical articles. Highlighting the model's performance through key metrics such as precision, recall, and F1 scores across various MeSH labels, this section lays the groundwork for a comprehensive analysis. We explore the implications of these results, identifying strengths and areas for improvement, and discuss the potential impact on future applications in biomedical text classification.

Test F1 Accuracy: 0.8557943679270112

	precision	recall	f1-score	support
A	0.78	0.86	0.82	2319
B	0.97	0.99	0.98	4662
C	0.89	0.88	0.88	2655
D	0.91	0.93	0.92	3083
E	0.84	0.92	0.88	3912
F	0.85	0.70	0.77	927
G	0.84	0.87	0.85	3329
H	0.55	0.18	0.27	562
I	0.72	0.62	0.67	581
J	0.71	0.58	0.64	549
L	0.67	0.50	0.57	755
M	0.88	0.90	0.89	2165
N	0.83	0.76	0.79	2304
Z	0.75	0.73	0.74	815
micro avg	0.86	0.85	0.86	28618
macro avg	0.80	0.74	0.76	28618
weighted avg	0.85	0.85	0.85	28618
samples avg	0.86	0.86	0.85	28618

Figure 2. Classification performance summary for DeBERTa-Enhanced multi-label classification

The results of our study as shown in Figure 2, underscore the model's efficacy in the intricate realm of multi-label classification within biomedical literature. The model achieved a Test F1 Accuracy of 85.58%, illustrating its capacity to balance precision and recall effectively across a wide array of MeSH labels. This achievement is indicative of the model's adeptness at comprehending and accurately categorizing biomedical articles, showcasing the potential of advanced NLP models in this domain.

An in-depth analysis of the classification report highlights variances in performance across different MeSH labels, reflecting the nuanced challenges inherent in multi-label classification tasks. For instance, the model demonstrated exceptional precision and recall in categories 'B' (97% precision, 99% recall) and 'D' (91% precision, 93% recall), evidencing its strong grasp of the semantic complexities within these areas. Conversely, the 'H' category, with a precision of 55% and recall of 18%, points to potential difficulties in classifying less represented or more complex topics, suggesting areas for future model enhancement.

Moreover, the micro-average scores, which aggregate the contributions of all classes to compute average metrics, showed an F1 score of 86%, precision of 86%, and recall of 85%. These micro-average metrics further validate the model's overall effectiveness in handling the extreme multi-label classification task, offering a holistic view of its performance across the entire label spectrum.

The variation in our model's performance across different categories emphasizes the need for ongoing enhancements to our approach. By focusing on strategies such as expanding the training dataset with additional examples for less represented labels, advancing our feature extraction methods, and incorporating cutting-edge NLP techniques, we aim to refine the model's ability to discern the subtle intricacies of biomedical literature more effectively. Concurrently, we recognize the necessity of addressing our study's limitations, including the potential biases within our dataset, the possibility of measurement errors, and the presence of confounding factors that may skew our results. A critical examination of our data collection methodologies, the representativeness of our sample size, and the influence of external variables is essential to bolster the credibility and generalizability of our research outcomes. This holistic approach to refining our model and rigorously assessing our research framework lays the groundwork for future investigations, ensuring that our contributions to the field of NLP in biomedical literature are both impactful and enduring.

In conclusion, our findings highlight the significant potential of deploying advanced transformer models, like DeBERTa, for the challenging task of extreme multi-label classification in biomedical literature. While the high F1 accuracy achieved signals the model's robust capabilities, the variation in performance across categories illuminates a path for future research and development. Our study makes a valuable contribution to the field of NLP applications in biomedical literature, providing insights into the strengths and areas for improvement within current methodologies.

V. CONCLUSION AND FUTURE WORK

In conclusion, our study represents a significant stride towards addressing the complexities inherent in the multi-label classification of biomedical literature. By leveraging the advanced capabilities of the DeBERTa model, we have demonstrated a robust methodology that not only achieves high accuracy in classifying articles into MeSH labels but also navigates the challenges of extreme label sparsity and imbalance. The detailed evaluation of our model's performance, as outlined in the "Results and Discussion" section, showcases its strengths in understanding and categorizing biomedical texts, offering a promising tool for enhancing information retrieval and knowledge discovery in the biomedical field.

However, the journey does not end here. Our findings have illuminated paths for future research, particularly in areas where the model's performance can be enhanced. Future work will focus on addressing the disparities in classification performance across different MeSH labels, potentially through the development of more sophisticated data augmentation techniques, advanced feature extraction methods, and the exploration of novel model architectures. Additionally, investigating the integration of domain-specific knowledge bases and ontologies could further refine the model's ability to understand and classify biomedical literature with even greater accuracy and nuance.

Moreover, the evolving landscape of natural language processing and machine learning presents an opportunity to explore new model variants and training strategies. The application of transfer learning, few-shot learning, and meta-learning approaches could offer novel pathways to overcoming the challenges of label imbalance and enhancing the generalizability of the model across diverse biomedical datasets.

In essence, our study lays a solid foundation for future explorations into the application of deep learning models for the classification of biomedical literature. By

continuing to innovate and refine our methodologies, we can unlock new potentials in biomedical research, accelerating the pace of discovery and contributing to the advancement of healthcare and medicine.

REFERENCES

[1] Q. Jin, R. Leaman, and Z. Lu, "PubMed and beyond: biomedical literature search in the age of artificial intelligence," eBioMedicine, vol. 100, p. 104988, Feb. 2024, doi: 10.1016/j.ebiom.2024.104988.

[2] "Medical Subject Headings - Home Page." https://www.nlm.nih.gov/mesh/meshhome.html

[3] J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org, Oct. 11, 2018. https://arxiv.org/abs/1810.04805

[4] A. S. Tejani, Y. S. Ng, Y. Xi, J. R. Fielding, T. G. Browning, and J. C. Rayan, "Performance of Multiple Pretrained BERT Models to Automate and Accelerate Data Annotation for Large Datasets," Radiology: Artificial Intelligence, vol. 4, no. 4, Jul. 2022, doi: 10.1148/ryai.220007.

[5] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, "Automated ICD coding using extreme multi-label long text transformer-based models," Artificial Intelligence in Medicine, vol. 144, p. 102662, Oct. 2023, doi: 10.1016/j.artmed.2023.102662.

[6] J. Xiong, L. Yu, X. Niu, and Y. Leng, "XRR: Extreme multi-label text classification with candidate retrieving and deep ranking," Information Sciences, vol. 622, pp. 115–132, Apr. 2023, doi: 10.1016/j.ins.2022.11.158.

[7] Q. Chen, J. Du, A. Allot, and Z. Lu, "LitMC-BERT: Transformer-Based Multi-Label Classification of Biomedical Literature With An Application on COVID-19 Literature Curation," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 5, pp. 2584–2595, Sep. 2022, doi: 10.1109/tcbb.2022.3173562.

[8] M. A. Ibrahim, M. U. Ghani Khan, F. Mehmood, M. N. Asim, and W. Mahmood, "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification," Journal of Biomedical Informatics, vol. 116, p. 103699, Apr. 2021, doi: 10.1016/j.jbi.2021.103699.

[9] A. F. Abdillah, C. B. P. Putra, A. Apriantoni, S. Juanita, and D. Purwitasari, "Ensemble-based Methods for Multi-label Classification on Biomedical Question-Answer Data," Journal of Information Systems Engineering and Business Intelligence, vol. 8, no. 1, pp. 42–50, Apr. 2022, doi: 10.20473/jisebi.8.1.42-50.

[10] "PubMed Multi Label Text Classification Dataset Processed.csv · owaiskha9654/PubMed_MultiLabel_Text_Classification_Dataset_MeSH at main." https://huggingface.co/datasets/owaiskha9654/PubMed_MultiLabel_Text_Classification_Dataset_MeSH/blob/main/PubMed%20Multi%20Label%20Text%20Classification%20Dataset%20Processed.csv.

[11] H. Kammoun, I. Gabsi, and I. Amous, "MeSH-Based Semantic Indexing Approach to Enhance Biomedical Information Retrieval," The Computer Journal, vol. 65, no. 3, pp. 516–536, Jul. 2020, doi: 10.1093/comjnl/bxaa073.

[12] G. Tsoumakas and I. Katakis, "Multi-Label Classification," International Journal of Data Warehousing and Mining, vol. 3, no. 3, pp. 1–13, Jul. 2007, doi: 10.4018/jdwm.2007070101.

[13] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 8, pp. 1819–1837, Aug. 2014, doi: 10.1109/tkde.2013.39.

[14] R. Alazaidah and F. Kabir, "Trending Challenges in Multi Label Classification," International Journal of Advanced Computer Science and Applications, vol. 7, no. 10, 2016, doi: 10.14569/ijacsa.2016.071017.

[15] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," Expert Systems with Applications, vol. 213, p. 118534, Mar. 2023, doi: 10.1016/j.eswa.2022.118534.

[16] P. He, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," arXiv.org, Jun. 05, 2020. https://arxiv.org/abs/2006.03654.

[17] Y. Jeong and E. Kim, "SciDeBERTa: Learning DeBERTa for Science Technology Documents and Fine-Tuning Information Extraction Tasks," IEEE Access, vol. 10, pp. 60805–60813, 2022, doi: 10.1109/access.2022.3180830.

[18] Z. Luo, "DecBERT: Enhancing the Language Understanding of BERT with Causal Attention Masks," arXiv.org, Apr. 19, 2022. https://arxiv.org/abs/2204.08688.

[19] "DeBERTa — DeBERTa 0.1.8 documentation." https://deberta.readthedocs.io/en/latest/modules/deberta.html.

[20] "DeBERTa." https://huggingface.co/docs/transformers/en/model_doc/deberta.

[21] J. Li, H. Tang, D. Tang and Z. Yang, "Multi-Label Zero-Shot Learning for Industrial Fault Diagnosis," 2023 6th International Conference on Information Communication and Signal Processing (ICICSP), Xi'an, China, 2023, pp. 1235-1240, doi: 10.1109/ICICSP59554.2023.10390617.

[22] I. Loshchilov, "Decoupled Weight Decay Regularization," arXiv.org, Nov. 14, 2017. https://arxiv.org/abs/1711.05101.

[23] I. Ofeidis, "An Overview of the Data-Loader Landscape: Comparative Performance Analysis," arXiv.org, Sep. 27, 2022. https://arxiv.org/abs/2209.13705.