Evaluation of Language Models for Multilabel Classification of Biomedical Texts

Panagiotis G. Syriopoulos, Andreas D. Andriopoulos,
and Dimitrios A. Koutsomitropoulos(✉)

Computer Engineering and Informatics Department, School of Engineering,
University of Patras, 26504 Patras, Greece
{st1059664,andriopa,koutsomi}@ceid.upatras.gr

Abstract. The continuous increase of data availability and the need for their utilization make it imperative to organize them into categories. Recent classification problems often involve the prediction of multiple labels simultaneously applying to a single instance. In this paper, we propose a structured approach for the implementation and evaluation of multilabel classification tasks in the context of biomedical texts. This involves selecting appropriate datasets and models, designing experiments, and defining metrics that accurately measure the models' performance across various aspects of the task. Our results yield notable scores and conclusions for the behavior of some state-of-the-art language models in specific data. It is shown that the complexity of biomedical data and the intricacy of multilabel classification require careful consideration of these models' capabilities to handle large label spaces, label correlations, and the nuances of biomedical language.

Keywords: multilabel classification · indexing · transformers · thesauri · MeSH · deep learning · biomedicine · PubMed

1 Introduction

Multilabel systems in biomedicine, with the integration of artificial intelligence and machine learning, display continuous evolution and expansion reflecting the complex nature of biological systems. One recent perspective is related to the use of transformer models [1], with studies rendering them capable of analyzing complex data structures by improving the accuracy of predictions [2]. Examples include the classification of clinical examinations, pathological conditions, diagnoses as well as multiple classification of medical texts [3]. These algorithms, with their ability to process big data and extract deep knowledge, offer significant potential for advancing medical science [4].

The intricacies of multilabel classification stem from the need to accurately assign multiple, often interrelated, labels to each instance, a task that becomes exponentially more challenging as the number of labels increases. This challenge is amplified in the biomedical domain where the accurate classification of data can directly impact the effectiveness of patient care, disease diagnosis, and the discovery of novel therapeutic

Published by Springer Nature Switzerland AG 2024
I. Maglogiannis et al. (Eds.): AIAI 2024, IFIP AICT 711, pp. 68–78, 2024.
https://doi.org/10.1007/978-3-031-63211-2_6

interventions [5]. Additional challenges posed by the biomedical context include the high dimensionality of data, the presence of noise and missing values, and the critical need for interpretable models that can be understood and trusted by practitioners [6].

Despite its significance, multilabel classification of biomedical data is a topic that has been relatively underexplored in the literature. This gap in research is partly due to the complexity of the task, which requires sophisticated algorithms capable of handling large, imbalanced datasets and the intricate correlations between labels [7].

In this paper we provide a comprehensive review of some state-of-the-art (SOTA) methods in multilabel text classification, with a particular focus on applications in the biomedical domain, including such language models as BioBERT [8], XLNet [9], DistilBERT [10], RoBERTa [11], ERNIE [12], and ELECTRA [13]. This review critically evaluates existing methodologies, highlighting their strengths, limitations, and suitability for various types of biomedical data. To our knowledge, there is no systematic review and evaluation of language models currently in the literature, that specifically addresses SOTA in multilabel classification of biomedical data with an increased number of labels.

Moreover, we propose a framework for multilabel classification that leverages the latest advancements in machine learning, including deep learning and transfer learning, to address the unique challenges of biomedical data. To this end, we use finetuning methods and employ metrics putting focus on already trained language models. Our source code is openly available at: https://github.com/pngsyr/Comparative-analysis-of-transformers-on-multi-label-pubmed-data.

The rest of this paper is organized as follows: in Sect. 2 we review relevant literature in the field of biomedical multilabel text classification; in Sect. 3 we summarize the models which we reviewed and evaluated; Sect. 4 presents our methodology and approach, by outlining the multilabel classification procedure designed, the underlying dataset and discusses optimizations regarding implementation, training and/or finetuning. Section 5 contains the results of the various experiments and their analysis, while Sect. 6 outlines our conclusions and future work.

2 Background and Related Work

The landscape of machine learning and its application in classification tasks within biomedical data has been the focal point of extensive research efforts over the past several decades. The emergence of sophisticated language models, particularly those based on deep learning architectures such as transformers [2], has revolutionized natural language processing (NLP). Adaptations of language models for biomedical tasks have shown promise, although primarily focusing on single-label classification problems or on those involving a relatively small number of labels [14].

Multilabel classification represents a significant leap in complexity from traditional single-label tasks. Most existing work, such as the ensemble method for multilabel classification [6] and the classifier chains model [15], has focused on scenarios with only a few labels. These foundational works have been crucial in understanding the intricacies of multilabel classification, including handling label correlations and imbalances. However, transitioning from single label to multilabel classification in this context is nontrivial, necessitating sophisticated approaches to accurately capture and predict multiple labels simultaneously [16]. Interdisciplinary approaches are also applied, as highlighted by recent reviews and studies on deep learning applications in medical imaging and genomics, and the detection of diseases such as COVID-19 using classification models [5].

There are also approaches for multilabel text classification with a large number of labels [17]. This study deals with the analysis of Inductive Conformal Prediction (ICP) for multilabel text classification and presents an approach to address the efficiency problem when dealing with a large number of distinct labels. By using LP (Label Powerset)-ICP and p-values, users can reject a large number of labels more easily and be guided to the appropriate results. Based on this knowledge, a classifier is created, which, based on semantic data, achieves better results than one that does not.

ML-Net [18] combines a label prediction network with an automatic mechanism for predicting the number of labels to provide an optimal set of labels. This is achieved by leveraging both the predicted confidence score of each label and the deep contextual information (modeled by ELMo) in the intended title. Specifically, ML-Net is evaluated on 3 independent text corpora in 2 types of text: biomedical literature and clinical notes. Evaluation metrics such as precision, recall, and F-measure are used for assessment.

Also, actions related to the activation of the researchers' engagement with methods of mining specific thematic categories from biomedical texts, such as that of COVID-19, are important. For example, authors in [19] organized the BioCreative LitCovid track to identify suitable topics in an automated way. The BioCreative LitCovid dataset, consisting of over 30,000 articles, was created for training and testing. It is one of the largest multilabel classification datasets in the biomedical scientific literature. The 19 participating teams managed to achieve a high F1 Score, up to 0.9394.

3 Models for Biomedical Classification

We have selected a range of transformer models for review and comparative evaluation. This includes general-purpose models like BERT, RoBERTa, and XLNet, alongside domain-specific models such as BioBERT, and models optimized for efficiency like DistilBERT and ELECTRA. Each model transforms text into numerical representations (embeddings) that encapsulate semantic and contextual information, crucial for managing the textual data's complexity. Below, a summary of each model and the rationale behind their effectiveness for multilabel classification problems is provided.

BioBERT is a domain-specific adaptation of BERT (Bidirectional Encoder Representations from Transformers) pretrained on large-scale biomedical corpora. It extends BERT's capabilities to biomedical texts, significantly improving performance on biomedical NLP tasks [8]. BioBERT's pretraining on biomedical literature (e.g., PubMed abstracts and PMC articles) makes it uniquely positioned to understand the complex jargon and nuanced meanings in biomedical texts. This enhanced understanding facilitates the accurate classification of documents into multiple, often closely related, biomedical categories.

XLNet is a generalized autoregressive pretraining model that outperforms BERT on various NLP benchmarks by capturing bidirectional contexts and overcoming limitations related to BERT's masked language model approach [9]. XLNet's ability to model the permutation of word sequences allows it to grasp the context better and understand the intricate relationships between terms in texts. This capability is particularly beneficial for multilabel classification, where understanding the interplay between multiple topics or labels within a single document is essential [20, 21].

DistilBERT is a distilled version of BERT that retains most of its predecessor's effectiveness while being 40% smaller and 60% faster. It demonstrates that much of BERT's performance can be preserved with significantly fewer parameters. Its efficiency and speed make DistilBERT ideal for multilabel classification tasks where computational resources are limited. Despite its reduced size, DistilBERT effectively captures contextual relationships within text, essential for accurately assigning multiple labels [10].

RoBERTa (Robustly optimized BERT approach) builds upon BERT's foundations with optimized training approaches, including dynamic masking, which leads to improved performance across a range of NLP benchmarks [22, 23]. RoBERTa's optimized training and better handling of context enable more nuanced understanding and prediction capabilities. This makes it highly effective for multilabel classification tasks, where distinguishing between subtle differences in text can be crucial for correct label assignment [11].

ERNIE (Enhanced Representation through Knowledge Integration) is designed to improve language understanding by incorporating world knowledge and entity information into pre-training. This approach allows ERNIE to better understand the semantics of words and phrases [24]. For multilabel classification, ERNIE's incorporation of external knowledge helps in accurately understanding and classifying texts that may require domain-specific insights or the recognition of nuanced relationships between entities, which is often the case in complex datasets [12].

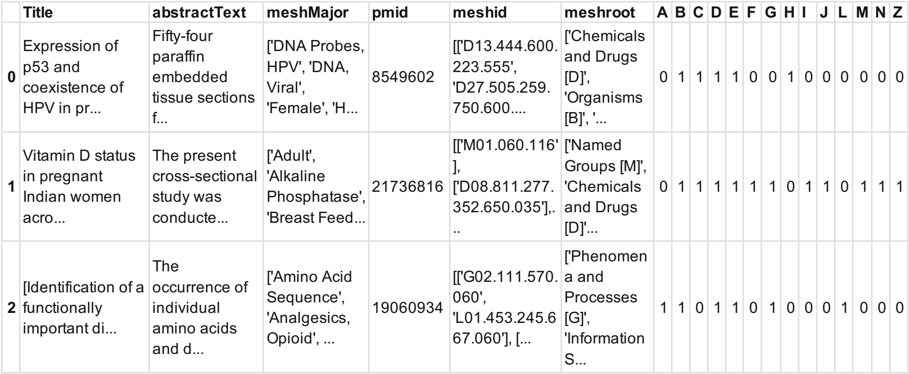
ELECTRA trains a more sample-efficient pre-training task called replaced token detection, rather than the masked language model used by BERT. It distinguishes between "real" and "fake" input tokens, which leads to more efficient learning. ELECTRA's efficiency in learning representations makes it suitable for multilabel classification, especially in scenarios where the dataset is vast or complex. Its unique approach to understanding context through the identification of token replacements allows it to accurately capture the nuances necessary for assigning multiple labels [13].

Each of these models brings distinct advantages to the task of multilabel classification, leveraging their architectural innovations and training methodologies to understand and process the complexities of language. Their effectiveness in biomedical contexts, particularly, underscores the importance of sophisticated NLP techniques in handling the multifaceted nature of biomedical data classification [25].

4 Methodology

4.1 Dataset

The dataset[1] consists of 50,000 biomedical articles from the PubMed library, which have been organized by domain experts for accurate annotation into 10–15 types of MeSH (Medical Subject Headings) labels. A sample of this dataset is depicted in Fig. 1.



	Title	abstractText	meshMajor	pmid	meshid	meshroot	A	B	C	D	E	F	G	H	I	J	L	M	N	Z
0	Expression of p53 and coexistence of HPV in pr...	Fifty-four paraffin embedded tissue sections f...	['DNA Probes, HPV', 'DNA, Viral', 'Female', 'H...	8549602	[['D13.444.600.223.555', 'D27.505.259.750.600....	['Chemicals and Drugs [D]', 'Organisms [B]', '...	0	1	1	1	1	0	0	1	0	0	0	0	0	0
1	Vitamin D status in pregnant Indian women acro...	The present cross-sectional study was conducte...	['Adult', 'Alkaline Phosphatase', 'Breast Feed...	21736816	[['M01.060.116'], ['D08.811.277.352.650.035'],...	['Named Groups [M]', 'Chemicals and Drugs [D]'...	0	1	1	1	1	1	1	0	1	1	0	1	1	1
2	[Identification of a functionally important di...	The occurrence of individual amino acids and d...	['Amino Acid Sequence', 'Analgesics, Opioid', ...	19060934	[['G02.111.570.060', 'L01.453.245.667.060'], [...	['Phenomena and Processes [G]', 'Information S...	1	1	0	1	1	0	1	0	0	0	1	0	0	0

Fig. 1. A sample of the PubMed dataset with label data.

Each entry in the dataset consists of the paper *Title*, the *abstractText*, which includes a summary of the topics mentioned in the specific paper, the *meshMajor*, which includes keywords related to the paper and help in searching based on these keywords, as well as the *pmid*, which is a unique identifier for each paper found in the PubMed library. It should be noted that the *MeSH IDs* are labels that are organized hierarchically in four levels, starting from the most general to the most specific ones. The first level corresponding to the most general topics (*mesh-roots*), consists of 14 labels, which correspond to the letters [‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘I’, ‘J’, ‘L’, ‘M’, ‘N’, ‘Z’]. The second, third, and fourth levels become increasingly specialized, focusing on a specific disease or condition. The mesh-roots, on the other hand, include the root tags in combination with the lexical representation of their category. The dataset has been preprocessed to flatten the deeper levels and map them to their respective roots. Therefore, there are the tags previously described, where each paper that includes a corresponding tag is marked with 1, otherwise, with 0 (one-hot encoding). In addition to this preprocessing, the dataset is shuffled and split 80/20 for training-finetuning and testing, respectively.

4.2 Implementation

The implementation environment used is Google Colab. The process begins with loading a pre-trained transformer model, whose weights have been trained on large text databases

[1] https://huggingface.co/datasets/owaiskha9654/PubMed_MultiLabel_Text_Classification_Dataset_MeSH.

and have already acquired a significant degree of language understanding. The dataset is transformed to align with the format required by each pretrained model. This includes the tokenization process, where the text is broken down into units (tokens), padding that enhances the data to have a uniform length, removing stop words, and encoding that converts the tokens into numeric values. A max_length of 128 has been specified for the tokenizer, with padding and truncation enabled, to ensure comparable finetuning of all pretrained models and align to the relative size of the abstracts contained in the dataset. Each example corresponds to multiple labels, and this requires specific management in the preparation of the labels. A summary of the implementation workflow is presented in Fig. 2.

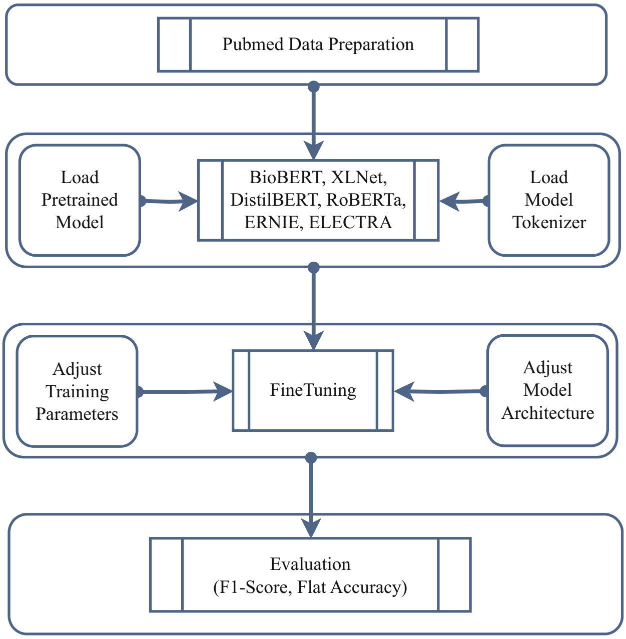

Pubmed Data Preparation
Load Pretrained Model
BioBERT, XLNet, DistilBERT, RoBERTa, ERNIE, ELECTRA
Load Model Tokenizer
Adjust Training Parameters
FineTuning
Adjust Model Architecture
Evaluation (F1-Score, Flat Accuracy)

Fig. 2. Implementation workflow.

The loss function used for training is *Binary Cross-Entropy*, as it can handle instances where each example belongs to more than one category. *Sigmoid* activation is applied to the model output to calculate probabilities for each label. In each training epoch, loss and gradients are calculated for various batches of data, and the model parameters are updated accordingly. Also, the model is evaluated on the validation data set. During validation, metrics such as the F1 Score and Flat Accuracy help to understand the model's performance (see Sect. 4.3).

The finetuning process is adopted to help preserve the knowledge the model has already acquired. In this approach, the model is trained in phases. Initially, layers that are added specifically for the downstream task of sequence classification are trained, while the rest of the layers of the pre-trained model remain frozen. This helps preserve the knowledge the model has already acquired. As training progresses, the frozen layers can gradually be unfrozen, allowing for further adaptation and improvement of the model to the task data. Finetuning occurs for a limited number of epochs (up to 12) on 80% of the dataset, allowing for further adaptation and improvement of the model to fit our specific task, while paying attention to preserving pretrained knowledge and monitoring critical performance metrics. As our focus is to assess the value of already pretrained models, we keep finetuning to as low as 12 epochs that are found to have a uniform impact on performance increase across all models.

4.3 Metrics

The F1 Score is the harmonic mean of precision and recall, calculated using the formula: F1 Score = 2 * (Precision * Recall)/(Precision + Recall). Precision measures the accuracy of the positive predictions made by the model, i.e., the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall measures the model's ability to identify all relevant instances, i.e., the proportion of correctly predicted positive instances out of all actual positive instances. It ranges from 0 to 1, where a higher value indicates better model performance in terms of both precision and recall. In the specific multilabel classification problem, we calculate precision, recall, and F1 Score for each label separately, and then compute the average F1 Score across all labels.

Flat Accuracy is a simpler measure of overall correctness which calculates the percentage of correctly classified instances out of all instances. In the context of multilabel classification, Flat Accuracy measures the proportion of instances where all predicted labels match exactly with the actual labels, regardless of the order or combination of labels. It provides a straightforward assessment of the model's performance in correctly predicting all labels for each instance.

To sum up, the F1 Score provides a balanced evaluation of precision and recall for each label separately, while Flat Accuracy provides a simple measure of overall correctness by considering all predicted labels collectively. Both metrics offer valuable insights into the performance of a multilabel classification model.

5 Evaluation

5.1 Configuration

For the purposes of evaluation, we use the validation dataset, which consists of 10,000 PubMed papers out of the total 50,000 papers, i.e. 20% of the original dataset. We evaluate the models on the validation dataset to assess their performance for the downstream task of sequence classification. We calculate F1 Score and Flat Accuracy based on the predictions generated by the model. We also utilize wandb.ai (Weights&Biases) library and environment for logging support and visualization of model training and evaluation metrics.

We load transformer models directly through HuggingFace using the Transformers library. In all cases we resort to the base variants for performance reasons.

We experiment with hyperparameter tuning, model architectures, and training strategies to optimize the model's performance further. For example, we use variable weight decay and employ the AdamW optimizer for training [26], an improvement over Adam that avoids the negative effect on the decay rate. We iterate on the training process based on insights gained from monitoring and analysis. All experiments are conducted on T4 GPU with 16GB of RAM made available by the Colab environment.

5.2 Results and Discussion

In the following, we report on the results of our experiments for the 6 language models surveyed. Figure 3 shows the values for Flat Accuracy with respect to the number of training epochs (1 to 12). Likewise, Fig. 4 contains the values for the F1 Score.

Overall, flat accuracy appears low, indicating the hardness of the problem but also the coarse and penalizing nature of the metric: Even if a single bit is wrong the whole sample is considered misclassified.

Most models show an increase in accuracy as the steps progress, which is to be expected, as the models learn from the data (Fig. 3). However, there are some fluctuations, which could be due to various factors such as learning rate adjustments or dataset problems. Models such as ERNIE, BioBERT and RoBERTa perform the best in terms of accuracy being twice as accurate as their counterparts and display higher peaks; this may indicate that they have the potential to achieve higher but may require more tuning or specific conditions to maintain peak performance.

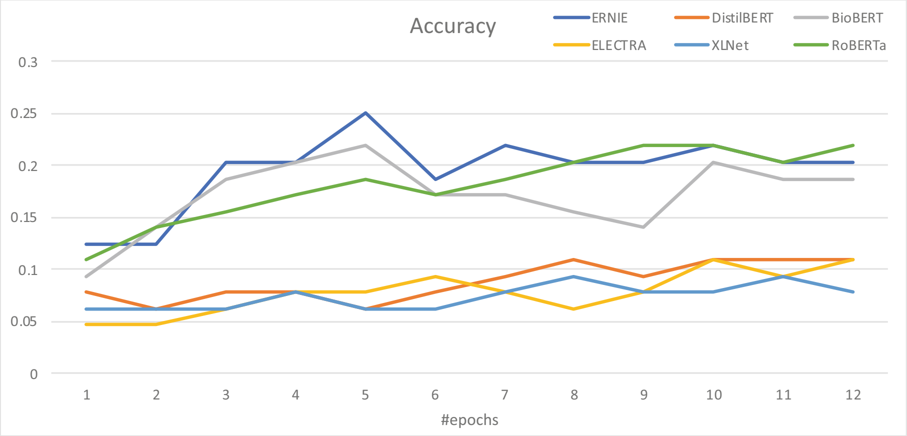


Fig. 3. Flat Accuracy results on the validation set wrt to training epochs.

In terms of the F1 Score, BioBERT appears to outperform all others, closely followed by RoBERTa and ERNIE (Fig. 4). It also exhibits consistently high precision and recall over almost all classes (not shown in the figure). This is possibly due to the model's adjustment and pretraining on domain-specific data i.e., biomedical texts. RoBERTa and ERNIE on the other hand exhibit comparably high F1 Scores, but their lack of specific pretraining lowers slightly precision and recall for some of the classes. Pretraining bias, if any, to the specific dataset tends to fade out however, after 6–7 finetuning cycles.

ELECTRA, XLNet and DistilBERT also perform well, but face challenges for specific classes and may require further finetuning and dataset balancing.

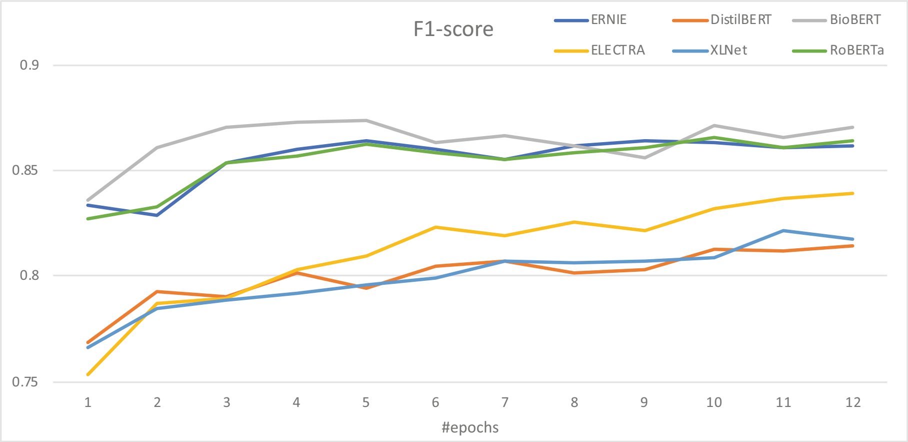


Fig. 4. F1 Score results on the validation set wrt to training epochs.

In Table 1, we present the overall results for train loss, F1 Score and Flat Accuracy on the validation set for each model.

Table 1. Training results

Models	Train loss	Val F1 Score	Val Flat Accuracy
ERNIE	0.2075	0.8618	0.2031
DistilBERT	0.3098	0.8139	0.1093
BioBERT	**0.1989**	**0.8700**	0.1875
ELECTRA	0.3411	0.8390	0.1093
XLNet	0.3381	0.8175	0.0781
RoBERTa	0.2566	0.8638	**0.2187**

As expected, BioBERT emerges as the top-performing model, closely followed by RoBERTa and ERNIE combining precision and accuracy. These models consistently show improvement in performance as finetuning progresses, indicating their capability to learn and adapt effectively to the data, regardless of their pretraining corpora. Also, the importance of dataset choice in achieving good results is highlighted. The appropriately sized dataset from PubMed, coupled with advanced transformer models, contributes significantly to the models' ability to generalize and perform well on multilabel classification tasks. The findings suggest that transformer models, when applied to biomedical text data like PubMed, have the potential to deliver high-quality results.

6 Conclusions and Future Work

Correctly predicting labels and tags for biomedical data is critical for various biomedical applications such as document classification, information retrieval, and knowledge extraction from scientific literature. At the same time, this task is far from trivial and can be costly, inefficient and error-prone. In view of the recent concerns and mishaps inflicting global health its importance is even more stressed. In this paper we have reviewed and evaluated a set of prominent language models that can alleviate the burden over the shoulders of experts. Our results suggest that transformer models demonstrate effectiveness in handling multilabel classification tasks on PubMed data; even more so when no domain-specific pretraining is required to produce adequate results, thus highlighting the versatility of transformer models in processing biomedical text.

Future research could extend to crafting hybrid models, merging for example ERNIE and BioBERT's strengths, leveraging BioBERT's biomedical expertise and ERNIE's adaptability across diverse data contexts. Delving into dimensionality reduction algorithms and data visualization techniques might yield deeper insights into biomedical article structures and limits of existing models, possibly leading to innovative pre-training strategies. Conducting experiments to assess model performance across various biomedical text types, such as clinical trials or review articles, could enhance model applicability. As biomedical informatics continues to evolve, continual evaluation and enhancement of machine learning models will be pivotal for field advancement.

References

1. Vaswani, A., et al.: Attention is all you need. In: Advances in neural information processing systems, pp. 6000–6010 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186 (2019)
3. Koutsomitropoulos, D.A., Andriopoulos, A.: Thesaurus-based word embeddings for automated biomedical literature classification. Neural Comput. Appl. (2021). https://doi.org/10.1007/s00521-021-06053-z.Springer
4. Wysocki, O., et al.: Transformers and the representation of biomedical background knowledge. Comput. Linguist. **49**(1), 73–115 (2023). https://doi.org/10.1162/coli_a_00462
5. Yoon, J., Kim, E., Yang, S., Park, S., Suh, J.-S.: A review of deep learning-based detection methods for COVID-19. Comput. Mater. Continua **65**(2), 1135–1152 (2019)
6. Martin, S.A., Townend, F.J., Barkhof, F., Cole, J.H.: Interpretable machine learning for dementia: a systematic review. Alzheimers Dement **19**(5), 2135–2149 (2023). https://doi.org/10.1002/alz.12948. (Epub 2023 Feb 3. PMID: 36735865)
7. U.S. National Library of Medicine. PubMed.gov. https://www.nlm.nih.gov/databases/download/pubmed_medline.html
8. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**, 1234–1240 (2019)
9. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Leet, Q.V.: XLNet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inform. Process. Syst. (2019)
10. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

11. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
12. Zhang, Z., Hun, X., Liu, Z., Jiang X.: ERNIE: Enhanced language representation with informative entities. In: Proceedings of the 57th annual meeting of the association for computational linguistics (2019)
13. Clark, K., Luong. M.T., Le, Q.V.D., Manninget, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
14. Kim, S., Lee, J., Gweon, G.: Deep learning in medical imaging: general overview. Korean J. Radiol. **21**(8), 945–958 (2020)
15. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333–359 (2011)
16. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., Telenti, A.: A primer on deep learning in genomics. Nat. Genet. **51**(1), 12–18 (2019)
17. Maltoudoglou, L., Paisios, A., Lenc, L., Martínek, J., Král, P., Papadopoulos, H.: Well-calibrated confidence measures for multi-label text classification with a large number of labels. Pattern Recognit. **122**, 108271 (2022)
18. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., Lu, Z.: ML-Net: multi-label classification of biomedical texts with deep neural networks. J. Am. Med. Inform. Assoc. **26**(11), 1279–1285 (2019). https://doi.org/10.1093/jamia/ocz085
19. Chen, Q., et al.: Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. Database (Oxford). (2022). https://doi.org/10.1093/database/baac069.PMID:36043400;PMCID:PMC9428574
20. Dai, Z., Yang, Z., Yang, Y., Jaime Carbonellet, Y.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Ilya Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
22. Liang, Y., et al.: XGLUE: a new benchmark dataset for cross-lingual pre-training, understanding and generation. arXiv preprint arXiv:2004.01401 (2020)
23. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP workshop blackbox NLP: analyzing and interpreting neural networks for NLP, pp 353–355. Brussels, Belgium, (2018)
24. Sun, Y., et al.: ERNIE 2.0: a continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34(05), pp. 8968–8975 (2020)
25. Houssein, E.H., Mohamed, R.E., Ali, A.A.: Machine learning techniques for biomedical natural language processing: a comprehensive review. IEEE Access **9**, 140628–140653 (2021). https://doi.org/10.1109/ACCESS.2021.3119621
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of international conference on learning representations (ICLR) 2019, arXiv preprint arXiv:1711.05101 (2019)