Simplifying Science: Utilizing BERT and SciBERT for Scientific and Plain Language Text Classification

Emily Robles
DATASCI 266: Natural Language Processing
UC Berkeley School of Information
emilyarobles@berkeley.edu

Abstract

Scientific jargon poses a significant barrier to the accessibility of scientific literature, yet from a researcher's perspective it can be difficult to identify . This study explores the efficacy of advanced natural language processing (NLP) models in distinguishing between scientific and plain language texts when fine-tuned using the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset. Leveraging the capabilities of BERT (Bidirectional Encoder Representations from Transformers) and SciBERT, a BERT variant pre-trained on scientific corpora, we conducted a comparative analysis to assess their performance in classifying a dataset as either scientific or plain language. Highlighting the advantages of domain-specific models in NLP tasks, SciBERT slightly outperformed BERT, although both achieved over 95% test accuracy and high F1 scores and recall. This research offers insights into the optimization of NLP models scientific text identification, which could lead to advancements in plain language tools to aid scientific communication.

Introduction

The effective communication of research findings is often hindered directly by the intricacy of scientific language. Scientific jargon, not only individual words but also the context and structure of their usage, often alienates a wider audience. In the medical research area for example, patients and their caregivers make up part of an audience that may lack the specialized scientific training needed to understand and interpret the findings of research papers but are nonetheless interested in and impacted by their findings.

Plain Language Summaries (PLS) coupled with scientific texts offer a solution to this issue and have been a large stepping stone towards democratizing science, however they can be difficult to write from a researcher's perspective. Our project assesses the effectiveness of employing NLP models to distinguish between "scientific" and "plain language" texts. Our goal is to set a foundation for tools to benefit scientific communicators by helping them identify language that may not be understood by a larger audience, particularly when creating Plain Language Summaries.

Background

The problem of overly technical language in scientific communication is well documented. One study of scientific abstracts published between 1881 and 2015 found that the readability is decreasing, and it is primarily due to the density of jargon[1]. This emphasizes the need for more comprehensible language in scientific literature at a time when NLP is well positioned to help. Tools like BERT have revolutionized text classification and language understanding, demonstrating remarkable capabilities in various linguistic tasks. The introduction of SciBERT by Beltagy et al. (2019) further extends these advancements for the scientific community, offering a model that is fine-tuned for scientific text and thus potentially more adept at handling the intricacies of scientific jargon[2].

There have been efforts to create tools to identify jargon text based on word frequency, such as the De-Jargonizer, [3] which has been used to analyze research abstracts and lay summaries and identify terminology differences between the two. A recently created but similar tool, MedJEx, was created by leveraging pre-trained models, including BERT and BioBERT[4], which reaffirmed our intuition to leverage BERT. Rather than focusing on individual words, however, our work intends to classify passages of text, primarily because the presence of appropriate context can make some "jargon" terms acceptable in terms of readability.

Methods

Data

The Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset contains 750 biomedical research paper abstracts, which were converted into plain language adaptations by medical informatics experts following annotation guidelines such as replacing arcane words like "orthosis" with common synonyms like "brace," changing sentence structure from passive voice to active voice, and explaining complex terms and abbreviations with explanatory clauses when first mentioned.[5]

We extracted the PLABA abstracts and adaptations for a binary classification task, where scientific text was labeled as '1' and plain language as '0'. The extracted dataset contained 1,671 text entries. We used an 80% training and 20% test split for the data, and did not include a validation set due to the limited size.

Preprocessing involved lowercasing and removing punctuation to reduce variability in the text, which was particularly important for our baseline model.

Scientific Text	**Plain Language Text**
The dystonias are a group of disorders characterized by excessive involuntary muscle contractions leading to abnormal postures and/or repetitive movements. A careful assessment of the clinical manifestations is helpful for identifying syndromic patterns that focus diagnostic testing on potential causes.	Dystonias are disorders with a lot of uncontrollable muscle contractions leading to awkward poses and/or repetitive movements. Checking the symptoms can help identify patterns that focus identification testing on possible causes.

Table 1. Excerpt from training data

Baseline Neural Network

Our baseline model, a simple neural network, utilized Word2Vec for embeddings and NLTK for text processing. This choice was driven by the model's simplicity and ease of implementation. However, we anticipated limitations in its ability to capture the nuances of scientific language given the model's rudimentary understanding of context and semantics.

BERT

We then progressed to using the base BERT (Bidirectional Encoder Representations from Transformers) for our first model. An advancement from our baseline, BERT's contextualized word embeddings capture the meaning of words based on their surrounding context which we believed would be critically important for differentiating scientific from plain language. Its bidirectional nature allows for a comprehensive understanding of text context, increasing its ability to classify text effectively. Additionally, when working with a dataset of limited size it is beneficial to work with a pre-trained model.

SciBERT

We then transitioned to exploring SciBERT, a variation of BERT fine-tuned on scientific text that we selected because of its training on a corpus of scientific papers. We expected it to outperform BERT because it is more attuned to the language and stylistic elements that occur in scientific literature.

Evaluation Metrics

We employed several key metrics to evaluate the performance of BERT and SciBERT models on the PLABA dataset. Accuracy was used to measure the overall proportion of correct predictions, indicating the model's general effectiveness in classification. The F1 Score provided a balanced measure of the model's accuracy in identifying true positives while minimizing false positives and negatives. Recall assessed the model's capability to correctly classify actual positive instances, thereby reflecting its effectiveness in reducing false negatives. Lastly, Test Loss quantified the model's prediction error rate on the test dataset, with lower values indicating more accurate predictions in line with the actual labels.

Results and Discussion

The baseline model yielded an accuracy of 53.13% on the test dataset. This performance underscores the model's limitations, particularly in grasping the contextual and semantic complexities of scientific language. The expectation is that BERT, with its advanced contextual understanding, will significantly outperform the baseline, demonstrating a more nuanced grasp of the language characteristics that distinguish scientific from plain text.

BERT (base)

The BERT base uncased model showed high performance and achieved a test accuracy of 97.01%, indicating its effectiveness in correctly identifying the majority of the text samples.

Evaluation Metrics

Test Loss: 0.0911
Test Accuracy: 97.01%
F1 Score: 0.9683
Recall: 0.9745

These metrics suggest a well-balanced model in precision and recall. The BERT model had six false positives and four false negatives in its predictions. The false positive examples were characterized by complex, technical language with dense medical and scientific terminology (e.g., "polyuriapolydipsia syndrome", "chronic fatigue syndrome", "fibroblast growth factor receptor 3"). We had expected to encounter more false positives than negatives because of the necessary inclusion of such technical language even within a plain language adaptation. Additionally, the texts incorrectly classified as scientific were lengthy and contain detailed explanations, which could have contributed to the model's confusion, potentially mistaking the depth of information for scientific content.

The false negative examples appeared to be more simplified or generalized discussions of medical or scientific topics (e.g., reports of myocarditis following COVID-19 vaccination, overview of newborn dried blood spot screening). These texts lacked the dense jargon and complex structures seen in the false positives and the language used seemed aimed at a broader audience, likely contributing to the model's misclassification as non-scientific.

SciBERT Model

We originally employed a simple iteration of the SciBERT model, but after reviewing the results from fitting on the training data became concerned about overfitting. The updated SciBERT model with dropout demonstrated similar results to BERT and slightly surpassed it in certain metrics. This suggests the efficacy of domain-specific training for tasks involving scientific text.

Evaluation Metrics
Test Loss: 0.0696
Test Accuracy: 97.91%
F1 Score: 0.9779
Recall: 0.9873

SciBERT had five false positive predictions and only two false negatives. Similar to BERT, SciBERT's false positives included texts with highly technical medical and scientific content (e.g., detailed descriptions of genetic conditions, treatment methodologies). These texts are specific in their subject matter and contain detailed, specialized information, likely leading to their misclassification as scientific. The false negatives again included texts that discuss scientific topics in a more general manner (e.g., a survey of myocarditis cases, an overview of newborn screening programs).

Conclusion

Both models tend to misclassify highly technical, jargon-heavy texts as scientific (false positives) and more generalized, accessible scientific discussions as plain language (false negatives). The length and complexity of the text, along with the density of specialized terminology, appear to be significant factors in misclassification. These findings suggest a potential area for improvement in model training, where a balance between technical jargon and general readability could be better calibrated. This analysis provides valuable insights into the types of texts that challenge BERT and SciBERT, highlighting the importance of context and language complexity in text classification tasks.

References

[1] Plavén-Sigray, Pontus, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. "The Readability of Scientific Texts Is Decreasing over Time." Edited by Stuart King. *eLife* 6 (September 5, 2017): e27725. https://doi.org/10.7554/eLife.27725.

[2] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." Edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), November 2019, 3615–20. https://doi.org/10.18653/v1/D19-1371.

[3] Rakedzon, Tzipora, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. "Automatic Jargon Identifier for Scientists Engaging with the Public and Science Communication Educators." *PLoS ONE* 12, no. 8 (August 9, 2017): e0181742.

https://doi.org/10.1371/journal.pone.0181742.

[4] Kwon, Sunjae, Zonghai Yao, Harmon S. Jordan, David A. Levy, Brian Corner, and Hong Yu. "MedJEx: A Medical Jargon Extraction Model with Wiki's Hyperlink Span and Contextualized Masked Language Model Score." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing 2022 (December 2022): 11733–51.

[5] Attal, Kush, Brian Ondov, and Dina Demner-Fushman. "A Dataset for Plain Language Adaptation of Biomedical Abstracts." Scientific Data 10, no. 1 (January 4, 2023): 8. https://doi.org/10.1038/s41597-022-01920-3.

Dormer, Laura, Thomas Schindler, Lauri Arnstein Williams, Dawn Lobban, Sheila Khawaja, Amanda Hunn, Daniela Luzuriaga Ubilla, Ify Sargeant, and Anne-Marie Hamoir. "A Practical 'How-To' Guide to Plain Language Summaries (PLS) of Peer-Reviewed Scientific Publications: Results of a Multi-Stakeholder Initiative Utilizing Co-Creation Methodology." *Research Involvement and Engagement* 8, no. 1 (June 2, 2022): 23. https://doi.org/10.1186/s40900-022-00358-6.

Stoll, Marlene, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. "Plain Language Summaries: A Systematic Review of Theory, Guidelines and Empirical Research." *PLOS ONE* 17, no. 6 (June 6, 2022): e0268789. https://doi.org/10.1371/journal.pone.0268789.