Classifying Texts into Organized Drug Label Sections Using BERT Language Modeling

U.S. FOOD & DRUG ADMINISTRATION
NATIONAL CENTER FOR TOXICOLOGICAL RESEARCH

Magnus Gray, Joshua Xu, Weida Tong, Leihong Wu
Division of Bioinformatics & Biostatistics
National Center for Toxicological Research, FDA
3900 NCTR Rd, Jefferson, AR 72079

Leihong Wu – Leihong.Wu@fda.hhs.gov

Introduction

- Structured data formats are critical for accessing and processing text data in regulatory science.
- Despite the FDA's guidelines for regulatory submissions in Structured Product Labeling (SPL) format[1], many labeling documents are still not well formatted, which hinders their information from being accessed and used.
- This study focuses on developing a language model that can classify texts/sentences into organized drug labeling sections.

Related Work

- Several studies have shown promising results with regard to predicting the section label for a given sentence.
- In 2017, Dernoncourt and Lee[2] aimed to predict the section (i.e., background, objective, method, result, or conclusion) of a given sentence for 200,000 medical abstracts, and their trained artificial neural network achieved an accuracy of over 90%.
- In 2020, researchers used BERT-based classification models to predict whether a sentence is propaganda (as well as the propaganda technique) or non-propaganda with 55-80%+ accuracy for various classification tasks.[3]

Methodology

- Over 17 million sentences were extracted from 45,626 drug labeling documents obtained from DailyMed's full release of human prescription labels.[4]
- Using Logical Observation Identifiers Names and Codes (LOINC)[5], the documents were separated into Physician Label Rule (PLR) Format (n=29,709) and Non-PLR Format (n=15,917).
- A series of BERT-based models were trained using datasets composed of 10,000 sentences per label from the "golden standard" PLR-format drug labeling documents.
- In addition to testing these models on PLR and Non-PLR datasets, an external dataset was assembled to further measure the models' performance.
- 9,580 Summaries of Product Characteristics (SmPCs), which are akin to U.S. prescription drug labels, were obtained from the UK medicine database, Electronic Medicines Compendium.[6]

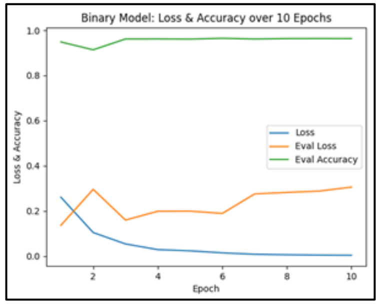
Results

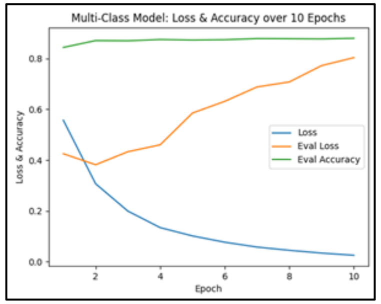
- The tables below show the prediction accuracies and precision produced by fine-tuned binary and multi-class models for three testing datasets, each of which contained 10,000 records per label.

		Overall	Indications & Usage	Warnings & Precautions
Binary *(96% Training Accuracy)*	*PLR*	**95%**	94%	96%
	Non-PLR	**89%**	91%	88%
	UK SmPC	**89%**	93%	85%

		Overall	Indications & Usage	Warnings & Precautions	Adverse Reactions	Other / Unknown
Multi-Class *(88% Training Accuracy)*	*PLR*	**84%**	90%	85%	86%	77%
	Non-PLR	**74%**	82%	47%	86%	80%
	UK SmPC	**67%**	79%	58%	71%	59%

- The results above show that the BERT-based model excels at differentiating sentences between two distinct categories and that model performance only slightly drops off when more categories are introduced.
- Furthermore, it is noted that the results for the Non-PLR and SmPC datasets are similar, showing that the model works well for the external UK dataset.





- The graphs above, showing the models' loss and accuracy over the course of fine-tuning, mostly display the expected behavior; however, the additional labels led to more evaluation loss in the multi-class model.

Discussion

- To explore how a different input-level would affect the results, the same experiments were conducted again, but with paragraphs as inputs rather than sentences.
- Typically, the paragraph-input models had better training accuracies and worse testing accuracies, indicating that overfitting was much greater in these models.
- Furthermore, the paragraph-input models were worse at sentence-label prediction than the sentence-input models.

Conclusion

- In an effort to make unstructured text information more accessible to researchers, this study focused on developing a language model that can classify texts into organized drug labeling sections.
- The results showed that automatically classifying free-texts into appropriate drug label sections is possible to an extent.
- Moreover, this research could be used to process other unformatted (e.g., scanned or photographed) drug labeling documents and their contents.

Acknowledgment

- Special thanks to ORISE – your support is appreciated.

References

1. U.S. Food & Drug Administration. Structured product labeling resources [Internet]. FDA; 2022 Apr 06. Available from: https://www.fda.gov/industry/fda-data-standards-advisory-board/structured-product-labeling-resources
2. Dernoncourt F, Lee JY. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071. 2017 Oct 17
3. Madabushi HT, Kochkina E, Castelle M. Cost-sensitive BERT for generalisable sentence classification with imbalanced data. arXiv preprint arXiv:2003.11563. 2020 Mar 16.
4. DailyMed. SPL resources: download all drug labels [Internet]. U.S. National Library of Medicine. National Institutes of Health; 2022 Mar. Available from: https://dailymed.nlm.nih.gov/dailymed/spl-resources-all-drug-labels.cfm
5. U.S. Food & Drug Administration. Section headings (LOINC) [Internet]. FDA. 2018 May 21. Available from: https://www.fda.gov/industry/structured-product-labeling-resources/section-headings-loinc
6. Datapharm. Latest Medicine Updates [Internet]. Electronic Medicines Compendium (EMC). Available from: https://www.medicines.org.uk/emc/