Mustering Machines' Meticulous Mastery: A Machine-Learning Model For Classifying Pharmacy Practice Publications Into Research Domains

Samuel O. Adeosun, RPh., PhD., Afua B. Faibille, BS., Aisha N. Qadir, Jerotich T. Mutwol, BS., and Taylor McMannen

Department of Clinical Sciences, Fred Wilson School of Pharmacy; High Point University, High Point, North Carolina

HIGHLIGHTS

- We proposed four research domains: Clinical, Education, Social & Administrative, and Basic & Translational, informed by the heterogeneity in pharmacy practice faculty publications, and the definitions of clinical research, clinical, and social pharmacy
- Our model (Pharmacy Practice Research Domain Classifier, **PPRDC**) had reproducible, state-of-the-art classification metrics and outperformed several general-purpose large language models including ChatGPT 4-o, Google Gemini 1.5 Pro, and Claude 3.5

BACKGROUND

- Although clinical and social pharmacy are essential elements of pharmacy practice, faculty publications are more heterogenous[1]
- Pharmacy practice is a low consensus field, where definition of terms vary among authors and across countries[1]
- Based on the Institute of Medicine's definition of Clinical Research[2], and the heterogeneity in pharmacy practice faculty publications, four research domains were proposed

OBJECTIVES

- To develop a deep neural network model (a subset of artificial intelligence and machine learning) for classifying pharmacy practice publications into the proposed domains, and to compare performance with state-of-the-art, general purpose large language models (gp-LLMs)

METHODS

- A **B**idirectional **E**ncoders **R**epresentations from **T**ransformer (**BERT**) model pretrained on biomedical corpus[3] was finetuned using 1000 random samples of abstracts from pharmacy practice faculty publications (publication years 2018-2021)
- The model was evaluated using 80 abstracts (publication year 2023) labelled with >80% consensus by all authors. Performance was compared with zero-shot performances of gp-LLMs like ChatGPT, Gemini etc.
- Metrics included F1, recall, precision and accuracy. Additional metrics included log-loss (a measure confidence of the model) and Cohen's kappa (a measure of agreement/reproducibility of predictions)
- A use case was demonstrated by testing the hypothesis that the pandemic did not affect domain distributions of publication before (2018-2019) and during (2020-201) the pandemic, using Chi-Square test of independence

RESULTS

Domain	F1	Recall	Precision	Accuracy
Clinical	92.7 (1.2)	93.2 (2.8)	92.2 (1.8)	92.4 (1.2)
Education	**96.2 (1.6)**	**97.6 (1.3)**	**94.9 (3.6)**	**98.7 (0.6)**
Social	85.8 (3.2)	83.4 (5.2)	88.9 (6.4)	92.7 (1.8)
Translational	83.1 (9.8)	86.5 (12.6)	80.1 (8.0)	98.2 (1.0)
Macro average	89.4 (1.7)	90.2 (2.2)	89.0 (1.7)	95.5 (0.6)

Table 1: Performance of the model on each research domain. Figures are mean (standard deviation) of 5-fold stratified cross-validation results of the model. **Domain with best metric in bold**; second best underlined.

Models	F1	Recall	Precision	Accuracy	Log loss	Cohen's K
PPRDC	**97.9**	**97.1**	**98.9**	**98.8**	**0.100**	**1.000**
ChatGPT 3.5	79.6	82.5	80.4	90.6	1.037	0.827
ChatGPT 4o	80.2	86.4	79.3	90.6	0.542	0.928
*Claude 3.5	91.1	90.6	93.8	95.6	0.267	0.978
Gemini 1.0	87.0	87.8	89.5	95.0	6.270	0.862
Gemini 1.5 Pro	92.6	95.3	91.3	96.9	2.257	0.941
LLAMA 3	84.7	86.0	86.5	93.1	1.669	0.959
Mistral Large	90.4	93.3	89.0	95.6	1.641	0.961

Table 2: Comparison of PPRDC with zero-shot performances of gp-LLMs. Figures are the best of 2 independent instances of zero-shot testing of each gp-LLM using the 80 abstracts labelled with ≥80% consensus by all authors. . Lower log-loss is better; higher Cohen's κ is better. **Model with best metric in bold**; second best underlined. *Sonnet

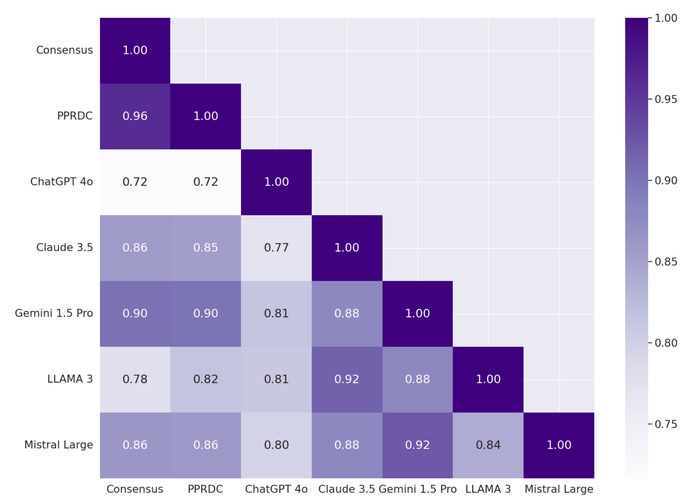

Figure 1: Pairwise Cohen's kappa among top 5 gp-LLMs & PPRDC. The best performance of each model was used; numbers represent Cohen's κ across the top-5 models (per F1 scores). Cohen's κ ranges from -1 to +1 (higher means better agreement). Test sample are the same 80 abstracts with ≥80% consensus labels as in the results in table 1.

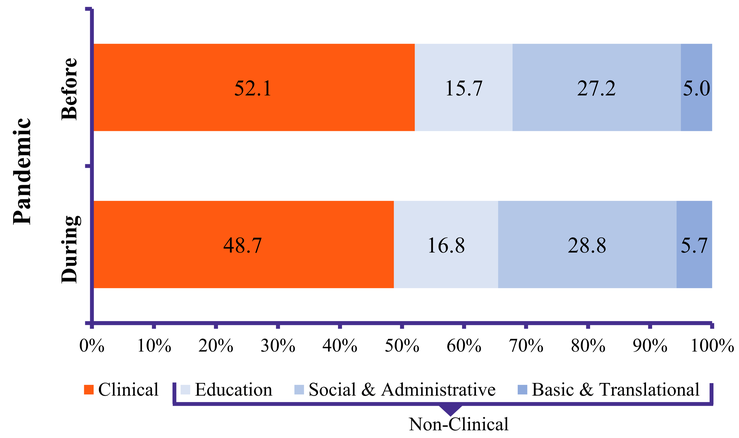

Figure 2: Use Case – Domain distribution of documents published before and during the pandemic. Chi-Square test showed no significant association when all four domains were considered in the analysis ($\chi2$=5.97, df=3, p=0.113); significant association when split into clinical & non-clinical domains ($\chi2$=5.67, df=1, p=0.017).

DISCUSSIONS

- PPRDC classification results are completely reproducible; it also produce probabilities of the abstract belonging to each of the four research domains
- The best and worst metrics were in education and social (respectively). This mirrors the contrasting heterogeneity and complexity within these research domains
- There were strong agreements among the gp-LLMs and PPPRDC

CONCLUSIONS

- **P**harmacy **P**ractice **R**esearch **D**omain **C**lassifier (**PPRDC**) accurately and reproducibly categorized abstracts into the four proposed research domains, and outperformed state-of-the-art general-purpose large language models in this task. This tool opens a new frontier in bibliometrics research and will facilitate consensus in pharmacy practice

REFERENCES

1. Fernandez-Llimos *et al*. Improving the quality of publications in advancing the paradigms of clinical and social pharmacy research: The Granada statements. (2023) *Res. Soc. Admin Pharm, 19 (5), 830-835.*
2. Tunis *et al*. Clinical Research Roundtable. Appendix V, Definitions of clinical research and components of the enterprise. (2002). https://www.ncbi.nlm.nih.gov/books/NBK220717/
3. Fang *et al*. Bioformer: An efficient transformer language model for biomedical text mining. (2023) *ArXiv*.

Pharmacy Practice Research Domain Classifier

Ask the presenter for a demo of the web-application at the poster session.