

**Reviewer's opinion
on Ph.D. dissertation authored by**

Klaudia Kantor

entitled:

*Neural language models for clinical
trial eligibility criteria*

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?

The central problem in this dissertation is investigating the degree to which free-text trial eligibility criteria are amenable to automation using neural models. This is not a trivial problem as the text may include temporal restrictions, exceptions, negations etc. Most current approaches are based on rule-based systems or (regular) expressions, which are insufficient and/or costly to automate, as demonstrated in the dissertation.

Is it a scientific one?

Yes. Although the hypothesis is straightforward (the use of LLMs to solving the automation of free-text trial eligibility criteria), the crux of the research relies on conducting an extensive review, performing a series of empirical experiments, critically assessing and interpreting the results, eliciting requirements for a software tool, and building a prototype tool. All these activities are part of the scientific endeavour. In addition, the chapters flow logically from beginning to end.

Does it have a practical meaning?

Certainly, this industrial PhD project is by design aimed at applicable results. Attesting to this are chapters 9 and 10 detailing the requirements for a suitable system and presenting a software tool. The tool is based, in a logical way, on the earlier chapters.

Contribution

What is the main, original contribution of the dissertation?

In my view there are three types of contributions: 1) Review/overview: The extensive scoping review with its good organization, and the overview of NLP resources in the biomedical domain, 2) The robust empirical studies providing useful insights into: the appropriateness of sentence embedding approaches; comparative investigation between GPT with BERT-based models; The advantages, but also challenges, of prompt engineering and 3) the elicitation of requirements for a practical tool and the presentation of a prototype system.

The dissertation strikes a good balance between the scientific and application aspects of the work. The quote "There is nothing more practical than a good theory" (from Lewin, 1952) applies to this dissertation.

2. Correctness

Can we trust what is claimed in the dissertation? Are the arguments correct? Indicate the flaws you have noticed, if any. Also point out those aspects concerning correctness that you value most (elegance of proofs, design of experiments, analysis of empirical data, quality of prototype software/hardware etc.).

The approaches adopted in this work are sound, so there are no major issues. Below are my comments on the accuracy of statements or design decisions that may have resulted in some improvements.

The dissertation states “The research hypothesis guiding this thesis is that neural language models can significantly enhance the efficacy of parsing clinical trial eligibility criteria, outperforming traditional methods *and consequently enhancing patient recruitment in clinical trials*” (my emphasis). The link between automatically processing the eligibility criteria to improving accrual is suggested, and theoretically this is certainly possible. However, it is likely that other recruitment factors are more important than interpreting the eligibility criteria. Quantifying the effect of automatically processing the eligibility criteria on accrual improvement is of course not crucial, as the research question of parsing the criteria in itself warrants research. Perhaps it is a good idea to explicitly state that coupling the parsing of eligibility criteria with a decision support system would lead into improvement in the process itself (the reminder or alert might be effective because of its specificity).

In chapter 3 the mechanism behind the missing values could have been considered (MCAR, MAR, MNAR). Imputing missing values with the mean is suboptimal and artificially lowers the variance in the dataset (using the chained equations would have been a better choice, as is done by the popular MICE framework).

In chapter 3 it is OK to consider values exceeding 2 standard deviations as “outliers” but being an outlier does not warrant exclusion. This approximately excludes 5% of the data and reduces the natural variability and can lead to bias in the analysis. Also the data may not be normally distributed so trimming based on two standard deviations assumes a symmetric distribution, which may not apply. Transformations or even Winsorization may have been more appropriate.

In chapter 3 the candidate made an interesting use of the c-index in a regression setting (I know this approach from survival analysis, but had not seen it before in a “normal” regression setting). The chapter states “A c-index of 1 indicates perfect prediction *accuracy*” The best term here is “*discrimination*” or “*concordance*”, not “*accuracy*”. A good thing about it is that it is unit-invariant. However, for the Mean absolute error it is better to either scale the values, or explicitly mention the unit in the tables.

Tables like 3.5 indeed provide insight. I would just note that one needs to be aware of the “Table 2 fallacy”, in which coefficients may be incorrectly interpreted in a causal way.

Chapter 5: very thorough overview. The noted limitation of not including arXiv is indeed notable, as in terms of methodology, it would have likely provided more interesting approaches.

Chapter 9: The requirements in chapter 9 could have been structured or semi-formalized like when using a requirements language such as MoSCoW.

3. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of **Information and Communication Technology**. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections? What is your opinion on the list of references? What is the degree of its completeness? Provide any other arguments in favour or against the claim that the candidate has general knowledge and understanding of the **Information and Communication Technology** discipline.

The candidate shows very good command of her field! The tutorial-like chapters include chapters 2, 3, 4 and 5 which are quite comprehensive. In addition, every chapter after that also includes background information, albeit obviously concise, demonstrating command of the material. The dissertation is inherently multidisciplinary and the candidate shows strong general knowledge in the following fields: medical research; biomedical resources; randomized trials with focus on expression of eligibility criteria; predictive modelling; NLP with focus on textual semantic similarity, Named Entity Recognition, prompt engineering; Requirements engineering; Software engineering.

In general, the grasp of the background knowledge is deep and the references are complete. The only domain I felt the background knowledge may be improved, admittedly only in some measure, is (preparations for) predictive modelling. Again, the demonstrated knowledge is certainly sufficient even in this domain.

4. Other remarks¹

I think the candidate could have claimed, at least explicitly, more use cases. Parsing eligibility criteria is obviously important in the context of patient accrual, but I feel there are other important use cases she could have claimed.

The first that comes to mind is to assist researchers for finding relevant trials (from <https://clinicaltrials.gov/> for example) based on specific eligibility criteria. In fact, one can think of a dynamic use case in which a trial designer may start typing eligibility criteria and a decision support system working in the background can start to suggest ("recommend") other criteria based on the given ones.

Another use case is using the results to help adapt the way we express our eligibility criteria in the first place. We often try to adapt technology to our needs, but sometimes we try to help the technology. I was wondering if companies like Roche could use the results of this thesis to improve the expression of their eligibility criteria in the trials (in order to help the system to interpret them correctly after that).

¹ Optional

My only remark in terms of the *scope* is that the dissertation's focus on breast cancer trials, as mentioned in the discussion chapter, may have been somewhat mitigated if we could try to characterize the RCTs used in breast cancer (what is generic in them? For example, what is the distribution of the elements such as negations, temporal restrictions, exceptions etc in them. This characterization could help assessing the generalizability of the approach.

I very much liked the proposed idea in future work in Chapter 7 of retrieval-augmented models for named entity recognition in biomedical texts. This I think holds a lot of promise.

Lastly, there are very minor issues like a typo in "the number of patient"; the layout of bullet points on page 26; the commas used instead of decimal points in the tables of Chapter 7 etc.

5. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the Act of 20 July 2018 - The Law on Higher Education and Science (with amendments)², my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with X)

☒

Definitely YES

☐

Rather yes

☐

Hard to say

☐

Rather no

☐

Definitely NO

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of **neural models/NLP for trial eligibility criteria**?

☒

Definitely YES

☐

Rather yes

☐

Hard to say

☐

Rather no

☐

Definitely NO

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

☒

Definitely YES

☐

Rather yes

☐


Hard to say

☐

Rather no

☐

Definitely NO


Signature

² <http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276>