Anca Dobrescu, Sarah Cohen-Boulakia and Nona Naderi

Attempt to rerun, reproduce and replicate Clinical Trials Sentence Classification Studies: lessons learnt

CONTEXT : Assessing Clinical Trial reporting with Natural Language Processing (NLP)

✅ Clinical Trials & Reporting Standards

Trials are essential for assessing medical interventions.
CONSORT guidelines [1] define what should be reported.

🚨 The Problem

Published articles often lack full compliance with CONSORT.
Leads to poor transparency and limited reproducibility.

🤖 NLP as a Solution

NLP models can automatically identify CONSORT items in text.
Need for annotated datasets and robust baseline models.

Use-Case Study

Study 1 [2]
CONSORT-TM corpus (10 709 sentences, 17 annotated items).
Methods: multi-label classification with SVM et BioBERT.

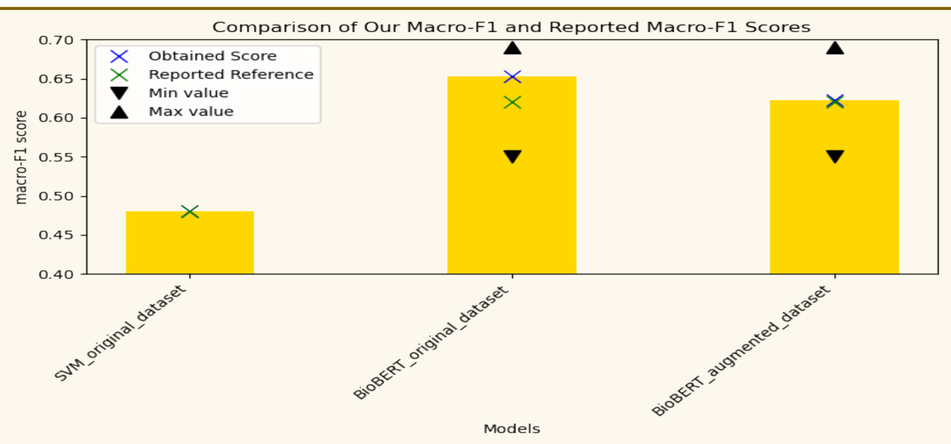
Study 2 [3]
Augmenting data from the CONSORT-TM corpus to improve detection of rare items.
Methods : UMLS and weakly supervision with Snorkel.

GitHub

Challenges

- The full CONSORT-TM corpus was available in XML, but preprocessing steps were undocumented or relied on unpublished in-house tools. Label distributions varied between articles and folds.
- Missing data: Class 0 (no CONSORT item) was used in experiments but never explained in the papers.
- Label inconsistencies: Some multi-labeled sentences retained only Method labels without justification; others with non-Method labels were reassigned to class 0.
- Evaluation challenges: GitHub repos lacked scripts for cross-validation and hyperparameter tuning. Details on metric computation (e.g., F1, AUC) and confidence intervals were not provided.

Results



Simply sharing articles, code, and data is not enough.
Rigorous alignment is essential.

Contribution : Alignment recommendations Article and Code Repository

Category	Article	Code repository
1.Data Collection	State the origin and describe the dataset (including class distribution).	Provide the dataset and mention its origin in the README.
2. Data Preprocessing	Detail all filtering and preprocessing steps, including any manual interventions.	Same as the paper, ideally with intermediate dataset versions.
3. Experimental Setup	Describe train/dev/test splits, class balancing, metric definitions, frameworks used, software/hardware environments, and model access.	Include all the above in the README, along with installation instructions and library versions.
4. Training Process	List hyperparameter search strategy and values, and the number of runs.	Provide the same in the README, with runnable code and commands.
5. Model Evaluation	Report results with central tendencies (e.g., mean, median), variation measures (e.g., standard deviation), and statistical tests.	Report results with central tendencies (e.g., mean, median), variation measures (e.g., standard deviation), and statistical tests.

Scan to rerun the studies.

Acknowledgments
ANR-22-CPJ1-0087-01
ANR-22-PESN-0007

[1] David Moher, Douglas G. Altman, Kenneth F. Schulz, and the CONSORT Group.2010. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMC Medicine 8, 1 (2010), 18. https://doi.org/10.1186/1741-7015-8-18.
[2] Halil Kilicoglu, Graciela Rosemblat, Linh Hoang, Sahil Wadhwa, Zeshan Peng, Mario Malički, Jodi Schneider, and Gerben ter Riet. 2021. Toward assessing clinical trial publications for reporting transparency. Journal of Biomedical Informatics 116 (2021), 103717. https://doi.org/10.1016/j.jbi.2021.103717
[3] L Hoang, L Jiang, and H Kilicoglu. 2022. Investigating the impact of weakly supervised data on text mining models of publication transparency: a case study on randomized controlled trials. AMIA Jt Summits Transl Sci Proc 2022 (May 2022), 254–263