MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230937

A Five-Step Workflow to Manually Annotate Unstructured Data into Training Dataset for Natural Language Processing

Yunshu ZHU[a], Ting SONG[a], Zhenyu ZHANG[a], Mengyang YIN[b] and Ping YU[a,1]
[a] *Centre for Digital Transformation, School of Computing and Information Technology, University of Wollongong, Wollongong, New South Wales, Australia*
[b] *Opal Healthcare, Sydney, Australia*
ORCiD ID: Yunshu Zhu https://orcid.org/0000-0003-2786-0775, Ting Song https://orcid.org/0000-0001-5858-6495, Zhenyu Zhang https://orcid.org/0000-0003-1853-4978, Mengyang Yin https://orcid.org/0000-0002-0212-4598, Ping Yu https://orcid.org/0000-0002-7910-9396

Abstract. Natural Language Processing (NLP) is a powerful technique for extracting valuable information from unstructured electronic health records (EHRs). However, a prerequisite for NLP is the availability of high-quality annotated datasets. To date, there is a lack of effective methods to guide the research effort of manually annotating unstructured datasets, which can hinder NLP performance. Therefore, this study develops a five-step workflow for manually annotating unstructured datasets, including (1) annotator training and familiarising with the text corpus, (2) vocabulary identification, (3) annotation schema development, (4) annotation execution, and (5) result validation. This framework was then applied to annotate agitation symptoms from the unstructured EHRs of 40 Australian residential aged care facilities. The annotated corpus achieved an accuracy rate of 96%. This suggests that our proposed annotation workflow can be used in manual data processing to develop annotated training corpus for developing NLP algorithms.

Keywords. Electronic health records, machine learning, annotation, annotation workflow, training data development, natural language processing

1. Introduction

Machine learning models have been powered by electronic health records (EHRs) [1], primarily relying on structured data from EHRs [2]. However, unstructured data, such as nursing notes, contain rich information that should not be ignored in machine learning [2,3]. Natural language processing (NLP) is required to extract useful information from unstructured data. However, it is challenging effectively utilise this data in NLP, which requires manual feature and label extraction, a time-consuming and labor-intensive process [4]. The availability of high-quality annotated text corpora [5-7] is critical for the performance of NLP algorithms.

Three methods for annotating text corpora are manual, semi-automatic, and automatic annotation [8]. Manual annotation, although the highest quality, often falls

[1] Corresponding Author: Professor Ping Yu. PhD, Centre for Digital Transformation, School of Computing and Information Technology, University of Wollongong, Wollongong, New South Wales, 2522, Australia, Phone: +61 2 4221 5412, email: ping@uow.edu.au.

below expectations due to a lack of effective guidance methods [9]. To address this, we propose a five-step workflow for manual annotation of unstructured EHRs.

2. Methods

We used a combination of literature review and expert-led, experience-based approach [10] to develop the five-step workflow for manual annotation of free-text data. We drew upon existing workflow and incorporated additional steps based on experts' experience to formulate the five-step workflow: (1) annotator training and reaching familiarity with the text corpus, (2) vocabulary identification, (3) annotation schema development, (4) annotation execution, and (5) result validation.

2.1. Step 1: Annotator training and familiarising with the text corpus

To ensure novice annotators understand the annotation purpose, it is beneficial to pair them with domain experts during the initial annotation task [10,11]. The domain experts can provide training, help them grasp the domain, and explain the different presentations of clinical terms. Annotators need to thoroughly review all data items and record relevant information for the specific topic. Domain experts can use the "think aloud" approach [12] to share their thoughts and comprehension with the annotators, facilitating their understanding of the data and its variations. This process continues until annotators gain confidence and familiarity with the data.

2.2. Step 2: Vocabulary identification

Once annotators are familiar with the text corpus, a coding dictionary is established as the annotation standard. It includes terms representing domain concepts with clear definitions [11]. The dictionary helps annotators determine which variables to annotate. It can be generated from existing ontologies like the Systematized Nomenclature of Medicine - Clinical Terms or manually created by domain experts analyzing EHRs.

2.3. Step 3: Annotation schema development

A standardized annotation schema ensures accurate and consistent annotation. It should encompass all relevant text in the EHRs and be applicable to the specific annotation project [9]. The schema is developed through two iterations. Initially, each annotator annotates a minimum of ten notes to create an initial schema, followed by discussions and consensus. In the second iteration, each annotator independently annotates 15 to 20 notes and compares results to reach agreement. The annotation process iterates with the updated schema until a complete and accurate annotated corpus is achieved. New terms not in the coding dictionary may be encountered during annotation. If deemed necessary, they should be added to the coding dictionary.

2.4. Step 4: Annotation execution

Annotators can use the final annotation schema to conduct annotation tasks, which can be recorded on a Microsoft Excel spreadsheet. The first column can be the patient

identifier. The second records the original text data. The rest columns record the label or features, which can be diseases, symptoms, or interventions, in binary format. Presence is annotated as "1", absence is "0". For example, if the original text is "He was refused to take medication" the corresponding agitation symptom is "refusal to care". Therefore, the target term is "refused to take medication", and the presence is "1".

2.5. Step 5: Result validation

To evaluate annotation performance, a 10% random sample of notes will be compared to the "ground truth" annotations by domain experts. Metrics such as precision, recall, F-score, and Kappa statistics will be used to measure accuracy and interrater reliability. These results will inform improvements to the annotation schema developed in Step 4.

3. Results

We conducted a case study to validate the effect of the five-step annotation workflow using the nursing progress notes of 4,445 residents from 40 residential aged care facilities (RACFs) in Australia. We followed the corresponding five steps to annotate the nursing progress notes to identify agitation symptoms. The study acquired ethics approval.

3.1. Step 1: Annotator training and familiarising with the text corpus

The annotator was trained by a registered nurse with more than ten-year work experience in Australian RACFs, and the two health informatics experts through six 2-hour Zoom meetings. They read ten notes of ten residents randomly extracted from the 40 RACFs to familiarise themselves with the content and discuss their understanding of the detailed information used to describe agitation symptoms (e.g., types of agitation).

3.2. Step 2: Vocabulary identification

An ontology, Dementia-Related Agitation Nonpharmacological Treatment Ontology [13], was selected as the base of the coding dictionary for annotation. It is the first comprehensive knowledge representation of nonpharmacological management for agitation in dementia, developed and evaluated by domain experts [13]. It summarised 67 agitation symptoms and their related definitions and synonyms. A three-hour Zoom meeting was conducted to discuss the content of each agitation symptom in ontology (e.g., the term definition and synonyms).

3.3. Step 3: Annotation schema development

We used an Excel spreadsheet to annotate the notes as it is easy to transfer data between different platforms. We randomly selected 55 notes from 2,024 notes (one note per resident) to develop the annotation corpus, using 40 notes to develop the annotation schema and 15 for validation.

3.4. Step 4: Annotation execution

We spent 334 hours developing the annotation schema using the DRANPTO coding dictionary. Our annotations used color coding, with red for agitation symptoms and green for "false positives." Out of 1,000 notes, 680 recorded dementia agitation symptoms. From these annotations, we identified 67 specific agitation symptoms to develop a rule-based NLP algorithm [14]. The F-score of the NLP algorithm is 89%.

3.5. Step 5: Result validation

We used Python programming language to randomly select 50 nursing notes for validation. The annotation error was used to refine the annotation schema. The accuracy of the annotation results reached 96%.

4. Discussion

We proposed workflow successfully annotated agitation symptoms in 680 out of 1,000 unstructured EHR notes, achieving a validated accuracy rate of 96% in a random sample of 50 annotations. Despite being labor-intensive and time-consuming, this five-step workflow can be applied to various annotation tasks, providing accurate training data for developing NLP algorithms. In our case study, we used the annotated corpus to develop a rule-based NLP algorithm specifically for extracting agitation symptoms [12].

The critical steps for achieving efficiency and quality of annotation are Step 1, annotator training and familiarising with the text corpus, and Step 3, annotation schema development. The importance of annotating the negative rules to reduce "false positive" cannot be undermined for the performance of the downstream NLP algorithm.

Our study had limitations due to resource constraints, leading to only one annotator performing the task. However, the annotator received training from three experienced domain experts to enhance accuracy. The time-consuming nature of the annotation task can cause annotator fatigue and increase error rates. To address this, the annotator took regular breaks at half-hour intervals to remain vigilant and reduce errors.

5. Conclusions

In response to the critical need to develop a high-quality machine learning data set, this study proposes a five-step workflow to manually annotating unstructured free-text EHR dataset. The usability of this workflow was demonstrated by applying it in a case study to annotate agitation symptoms from the unstructured EHRs, which reached a 96% accuracy rate. Therefore, our five-step workflow can be used to guide the manual annotation of free-text corpus for the downstream NLP tasks.

Acknowledgements

The authors would like to acknowledge the aged care organisation that shared the electronic health records with us and made this study possible.

References

[1] Bayramli I, Castro V, Barak-Corren Y, Madsen EM, Nock MK, Smoller JW, Reis BY. Predictive structured-unstructured interactions in EHR models: A case study of suicide prediction. NPJ Digit Med. 2022 Jan;5(1):15, doi: 10.1038/s41746-022-00558-0.

[2] Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019 Apr;26(4):364-79, doi: 10.1093/jamia/ocy173.

[3] Park J, You SC, Jeong E, Weng C, Park D, Roh J, Lee DY, Cheong JY, Choi JW, Kang M, Park RW. A Framework (SOCRATex) for Hierarchical Annotation of Unstructured Electronic Health Records and Integration Into a Standardized Medical Database: Development and Usability Study. JMIR Med Inform. 2021 Mar;9(3):e23983, doi: 10.2196/23983.

[4] Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. Curr Epidemiol Rep. 2018 Dec;5(4):331-42, doi: 10.1007/s40471-018-0165-9.

[5] Li X, Ma D, Yin B. Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system. Comput Electron Agric. 2021 Jan;180:105908, doi: 10.1016/j.compag.2020.105908.

[6] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49, doi: 10.1016/j.jbi.2017.11.011.

[7] Kotecha D, Asselbergs FW, Achenbach S, Anker SD, Atar D, Baigent C, Banerjee A, Beger B, Brobert G, Casadei B, Ceccarelli C, Cowie MR, Crea F, Cronin M, Denaxas S, Derix A, Fitzsimons D, Fredriksson M, Gale CP, Gkoutos GV, Goettsch W, Hemingway H, Ingvar M, Jonas A, Kazmierski R, Løgstrup S, Thomas Lumbers R, Lüscher TF, McGreavy P, Piña IL, Roessig L, Steinbeisser C, Sundgren M, Tyl B, van Thiel G, van Bochove K, Vardas PE, Villanueva T, Vrana M, Weber W, Weidinger F, Windecker S, Wood A, Grobbee DE; Innovative Medicines Initiative BigData@Heart Consortium, European Society of Cardiology, CODE-EHR international consensus group. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. Eur Heart J. 2022 Oct;43(37):3578-88, doi: 10.1093/eurheartj/ehac426.

[8] Wei Q, Franklin A, Cohen T, Xu H. Clinical text annotation - what factors are associated with the cost of time? AMIA Annu Symp Proc. 2018 Dec;2018:1552-1560.

[9] Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform. 2006 Apr;39(2):196-208, doi: 10.1016/j.jbi.2005.06.004.

[10] Song T, Yu P, Bliokas V, Probst Y, Peoples GE, Qian S, Houston L, Perez P, Amirghasemi M, Cui T, Hitige NPR, Smith NA. A Clinician-Led, Experience-Based Co-Design Approach for Developing mHealth Services to Support the Patient Self-management of Chronic Conditions: Development Study and Design Case. JMIR Mhealth Uhealth. 2021 Jul;9(7):e20650, doi: 10.2196/20650.

[11] Yordanova K, Krüger F. Creating and Exploring Semantic Annotation for Behaviour Analysis. Sensors (Basel). 2018 Aug;18(9):2778, doi: 10.3390/s18092778.

[12] Altalhi F, Altalhi A, Magliah Z, Abushal Z, Althaqafi A, Falemban A, Cheema E, Dehele I, Ali M. Development and evaluation of clinical reasoning using 'think aloud' approach in pharmacy undergraduates - A mixed-methods study. Saudi Pharm J. 2021 Nov;29(11):1250-7, doi: 10.1016/j.jsps.2021.10.003.

[13] Zhang Z, Yu P, Chang HCR, Lau SK, Tao C, Wang N, Yin M, Deng C. Developing an ontology for representing the domain knowledge specific to non-pharmacological treatment for agitation in dementia. Alzheimers Dement (N Y). 2020 Sep;6(1):e12061, doi: 10.1002/trc2.12061.

[14] Zhu Y, Song T, Zhang Z, Deng C, Alkhalaf M, Li W, Yin M, Chang HCR, Yu P. Agitation Prevalence in People With Dementia in Australian Residential Aged Care Facilities: Findings From Machine Learning of Electronic Health Records. J Gerontol Nurs. 2022 Apr;48(4):57-64, doi: 10.3928/00989134-20220309-01.