Efficient Elicitation of Fictitious Nursing Notes from Volunteer Healthcare Professionals

Jesper Vaaben Bornerup
IT University of Copenhagen
jesper.bornerup@live.dk

Christian Hardmeier
IT University of Copenhagen
chrha@itu.dk

Abstract

Reliable automatic solutions to extract structured information from free-text nursing notes could bring important efficiency gains in healthcare, but their development is hampered by the sensitivity and limited availability of example data. We describe a method for eliciting fictitious nursing documentation and associated structured documentation from volunteers and a resulting dataset of 397 Danish notes collected and annotated through a custom web application from 98 participating nurses. After some manual refinement, we obtained a high-quality dataset containing nurse notes with relevant entities identified. We describe the implementation and limitations of our approach as well as initial experiments in a named entity tagging setup.

1 Introduction

With the emergence of Electronic Health Records (EHR), the way nurses document their work has changed drastically. Printed schemas and handwritten notes were supplanted by computer-based systems like the Danish Sundhedsplatformen (SP), aiming to reduce data redundancy and errors (Ambinder, 2005). To simplify automatic processing and data reuse, EHR systems emphasize structured documentation. This choice has been described as “Technological somnambulism” (Johnson, 2016) and tends to be at odds with the preferences of the clinical professionals, who value usability and flexibility (Rosenbloom et al., 2011) and experience structured documentation as time-consuming and inefficient (Brinkmann et al., 2020; Baumann et al., 2018), frequently leading to inadequate documentation (Tram, 2017).

Automatic generation of structured documentation from free-text nurse notes would offer an attractive solution to this dilemma. However, the development of such systems across countries and languages is frustrated by the lack of training data due to the stringent privacy constraints surrounding all forms of medical notes (Landolsi et al., 2023). While some relevant datasets are available (Johnson et al., 2016), they are specific to the context in which they were produced and may be of limited use in another location characterised by a different language, different social context or different healthcare procedures.

In this paper, we describe and evaluate a method to elicit fictitious nurse notes from volunteering healthcare professionals based on visual stimuli. The collected notes closely mirror real free-text nursing documentation without suffering from the privacy restrictions of authentic notes. Emphasising a low time commitment for the volunteers, our method enabled us to collect a high-quality dataset of 397 notes from 98 participating nurses. We describe our procedures for eliciting and curating the dataset and annotating it for information extraction as well as initial experiments on automatic extraction of structured data. Our dataset is in Danish, but the procedure would be easily generalisable to other languages.

2 Data collection framework

We collected fictitious examples of nursing notes, together with structured annotations of their content, with two goals in mind: 1. The notes collected should mimic authentic nursing notes as much as possible. 2. The entry threshold for participants should be minimal to make recruitment easier. We used visual stimuli to minimize the influence of the stimuli on the participants’ word choice, and imposed a time limit on the text entry to simulate real-life time pressure.

Figure 1 shows the structure of our web application, whose core parts are the stimulus presentation, note capture and structured annotation. Dif-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 739–754
March 3-4, 2025 ©2025 University of Tartu Library

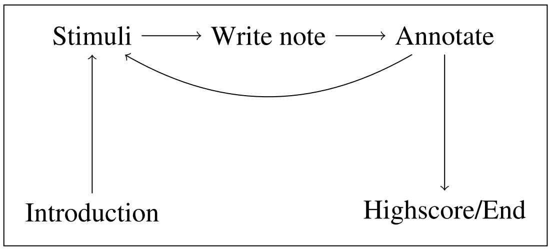
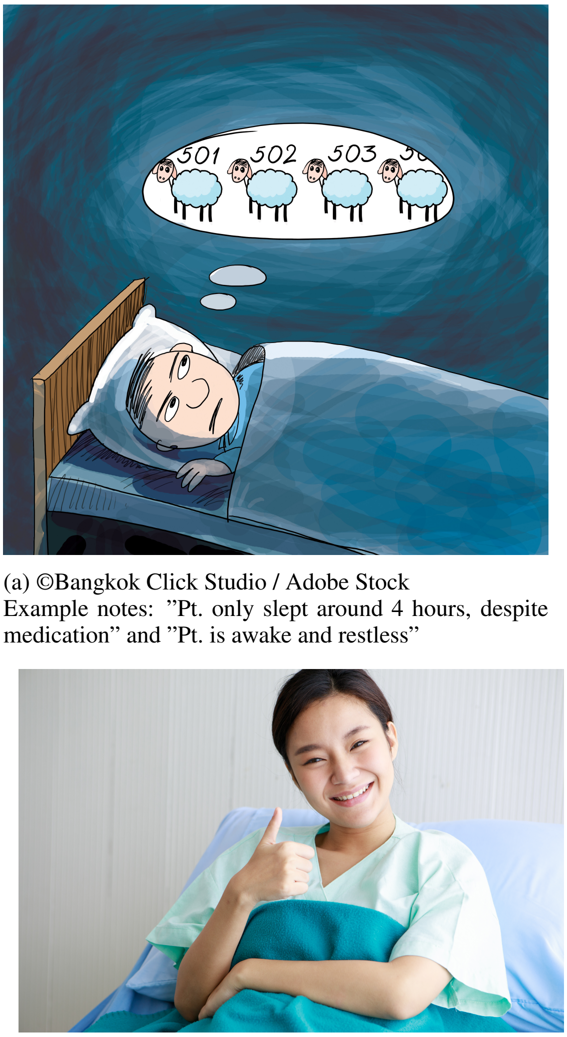


Figure 1: Data Collection Process. After annotating the participant gets the option to repeat or stop.

ferent sets of test participants were used to evaluate the design and offer feedback on the web application during the design process. Some of the test participants were observed doing the process, other were interviewed afterwards.



(a) ©Bangkok Click Studio / Adobe Stock
Example notes: ”Pt. only slept around 4 hours, despite medication” and ”Pt. is awake and restless”

(b) ©Andrius Gruzdaitis / Adobe Stock
Example notes: ”Pt. feeling better and is ready to get discharged later today” and ”Pt. happy with the plan and will contact the department in case of worsening in symptoms”

Figure 2: Stimuli examples

As we considered a denser and more focused dataset more useful than a sparse dataset covering many areas, some of the nurse-relevant problem areas were omitted in the our data collection to increase the number of items per category.

Introduction. The introduction page consists of a 4-step guide, including three small video clips demonstrating the process of seeing a stimulus, writing a note and annotating it.

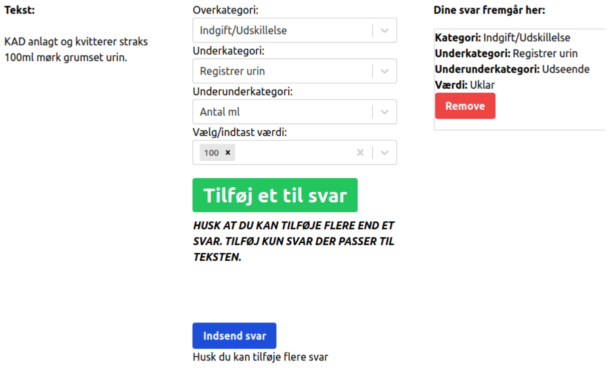
Initially the introduction included detailed instructions to the participants. However, during testing, most of the test participants did not read the text and quickly pressed ”next” to move on to the next step, which led to confusion about the process. To mitigate this, the text was cut significantly and the introduction page was redesigned with three GIF animations demonstrating the process. The Facebook post advertising this study also described that the purpose was to create fictitious free-text nursing documentation.

Stimulus presentation. The stimulus display page features an image or video, a 60-second countdown timer and a button to manually progress. The stimulus is drawn uniformly at random from 23 unique items (16 pictures, 7 videos), each chosen to inspire the participants to write relevant nursing documentation. Examples of stimuli and associated notes are shown in Figure 2.

Note capture. The write note page consists of 6 fields in which the participants can write notes based on the 12 nursing-related problem areas (*sygeplejefaglige problemområder*) defined by Styrelsen for Patientsikkerhed (Danish Patient Safety Authority) (Styrelsen for Patientsikkerhed (SFPS), 2023), which defines minimum requirements for nursing documentation. Given the anticipated limited volume of collected data, certain problem areas, including pain and sexuality, were excluded to ensure a more targeted dataset.

A time limit, randomly selected in 9 steps from 20–135 seconds, was imposed on the participants.

Structured annotation. The structured annotation page, shown in Figure 3, is composed of three sections. On the left, the note intended for annotation is displayed for the participant. The right section presents the completed annotations, while the central area houses the module responsible for managing the annotation process. The design of this system adopts a similar layered struc-



Tekst:

KAD anlagt og kvitterer straks 100ml mørk grumset urin.

Overkategori: Indgift/Udskillelse

Underkategori: Registrer urin

Underunderkategori: Antal ml

Vælg/indtast værdi: 100

HUSK AT DU KAN TILFØJE FLERE END ET SVAR. TILFØJ KUN SVAR DER PASSER TIL TEKSTEN.

Husk du kan tilføje flere svar

Dine svar fremgår her:

Kategori: Indgift/Udskillelse
Underkategori: Registrer urin
Underunderkategori: Udseende
Værdi: Uklar

Figure 3: Annotate page

ture found in the schemes of EHRs, with a categories, subcategories and subsubcategories to narrow down the options for the final selected value. There is a one-to-one relationship between the highest-level categories and the six fields in the note capture section.

Highscore. The highscore page showed the top contributors and gave the participants a choice to end the process or take one more cycle.

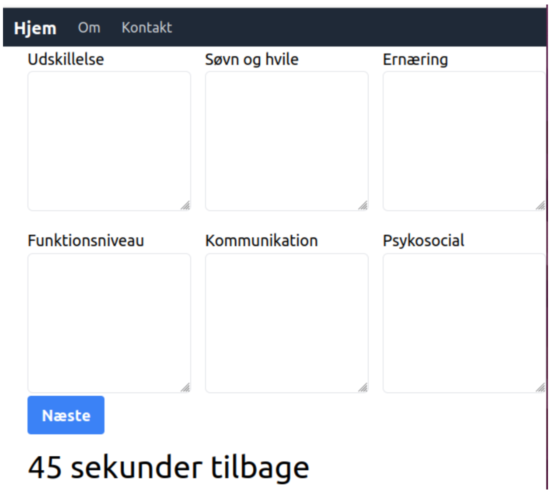
3 Collected data

The study was advertised four times in a Facebook group with 30,000 nurses, and three medical wards were visited once each to recruit participants. A total of 98 nurses participated in the study, producing 407 notes and 594 annotations. We expect that this number could be increased by offering economic incentives for participation. Every note and annotations was manually reviewed for quality control.

3.1 Notes

Most participants produced 1 note (n=34), and the average number of notes per person is 3.75. Typical notes are short and concise with an average length of about 8 words per note, focusing on one category per note. 16 out of 407 notes (3.9%) had to be removed, because they had a length of 1 word, because they directly described the stimulus shown or because they were spam.

The length of the notes shows a very slight upward trend as the time limit was increased, but the effect is not very strong (Figure 5). This might be attributed to participants having the option to proceed by clicking “next” at their discretion, before the timer ran out.



45 sekunder tilbage

Figure 4: Note capture page

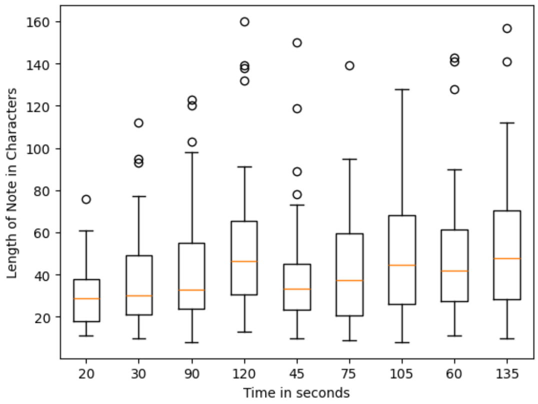


Figure 5: Average Note Length per Timer

3.2 Structured annotations

Each annotation consists of a category, a subcategory, a subsubcategory and a value. Figure 6 shows a note with 4 annotations. The subcategory is not only used to navigate to the right subsubcategory, it also carries information that relates to the final value.

Unannotated. 64 of the notes were submitted without any annotations. 14 were impossible to annotate, as there was no type of annotation which would fit the note, 12 were either 1 character long or cut short, probably because of the time limit, and 38 were possible to annotate.

Annotated. The remaining 343 notes had annotations. The annotations can be divided into 4 groups, all represented in Figure 6.

1. **Exact match:** The selected value in the annotation has an exact match in the note.

Annotation Type	Count	Percentage
Total annotations	594	100.0%
Exact match	297	50.0%
Partial match	106	17.8%
Interpretation	78	13.2%
Incorrect irrelevant	39	6.5%
Incorrect relevant	74	12.5%

Table 1: Annotation Statistics

2. **Partial match:** The selected value in the annotation has partial overlap with the note. This could happen because of two reasons.
   (a) The choices offered by the annotation process forced the use of another word, than was in the note. The structured part enforces the use of the Bristol Stool Scale (Lewis and Heaton, 1997) (which defines consistencies of stools) where "type 4" amounts to "soft".
   (b) The entity in the note was misspelled or in plural form, causing a mismatch with the structured category.
3. **Interpretation/classification:** The selected value can be interpreted by the note. In Figure 6, the amount of persons is not mentioned, however operating a ceiling hoist requires two people, making the annotation correct.
4. **Incorrect:** The annotation fits in none of the above categories. These annotations can be divided into two categories:
   (a) Relevant, where the annotation fits the theme, but is not present in the note. In Figure 6 the size of the stool is annotated, but is not present in the note.
   (b) Irrelevant, where the annotation is completely unrelated.

Missing Annotations. Missing annotations occur when an *Exact match* or *Partial match* annotation is possible, but missing. Omitted possible *Interpretation* annotations are not considered missing due to the subjectivity of this category. A total of 107 annotations were missing. The distribution among the types of annotation can be seen in Table 1.

A total of 64 different *subcategory/subsubcategory* pairs were used by the participants, with the 5 highest having from 22 to 55 entries and the lowest 5 having one entry each.

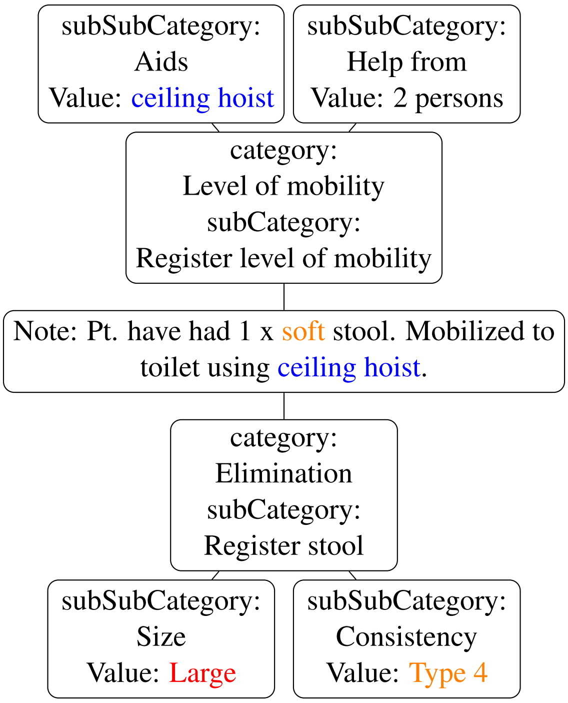

subSubCategory: Aids
Value: ceiling hoist
subSubCategory: Help from
Value: 2 persons
category: Level of mobility
subCategory: Register level of mobility
Note: Pt. have had 1 x soft stool. Mobilized to toilet using ceiling hoist.
category: Elimination
subCategory: Register stool
subSubCategory: Size
Value: Large
subSubCategory: Consistency
Value: Type 4

Figure 6: Top annotations: Left Exact match, right Interpretation
Bottom annotations: Left Incorrect, right Partial match .

3.3 Data evaluation

Four people replied to the Facebook post advertising the study that they did not understand the task, and another wrote the interface was too confusing. No other feedback from participants was received.

3.3.1 Notes

A manual review of the notes showed good variety in word choice (e.g., 'murky' and 'unclear' used interchangeably) and a realistic feel, suggesting they could have been real nurse documentation. The goal was to balance stimuli uniformly across the 6 main categories, but the resulting dataset is not balanced (Figure 2). This could be because some stimuli were harder to understand and therefore harder to write a note to or because some stimuli could be interpreted in multiple ways. For example, a picture of a diaper could both represent *elimination* and *mobility*.

3.3.2 Annotations

The structured annotation part posed a greater challenge for the participants, resulting in 64 unannotated notes (18.6%). However, 12 of those

Category	Count	Percentage
Elimination	145	24.4%
Mobility	133	22.4%
Psychological and social	83	14.0%
Sleep and rest	81	13.6%
Communication	78	13.1%
Nutrition	74	12.5%

Table 2: Category distribution

were errors or probably cut short because of the time limit, which can be expected. 14 were impossible to annotate with the options given to the participants. This leaves 38 (11%) of the notes which were possible to annotate, but had no annotations.

Incorrect annotations amount to 19% of all annotations, with 66% of them relevant to the topic and the rest completely irrelevant. These were removed from the dataset.

Missing annotations also pose a significant problem. Missing annotations and unannotated notes may be due the interface of the annotation process. While the interface mimics a real EHR, it is not exactly the same. They may also reflect the restrictions of structured documentation: It is time consuming, and finding the right category can be difficult (Brinkmann et al., 2020; Baumann et al., 2018). With no tangible incentive to spend time on it, participants may just click next and move on if they cannot find the right category immediately.

Users had the ability to add their own entity, if it was not among the options provided by the web application. This was however not utilized and that could be the reason for some of the missing annotations.

64 distinct subCategory/subSubCategory pairs were utilized by participants, with the majority being used less than 8 times. This posed a significant challenge for the experimental part of our study (extracting structured information from free-text nurse documentation). To simplify the problem, the classification part of the annotations was discarded as they represented a very small part of the annotation. The remaining annotations were either an exact or partial match, enabling us to reframe the task as a Named Entity Recognition (NER) challenge. Here, the subSubCategory represents the *entity type*, while the value represents an *instance* of the entity type.

4 Entity tagging

Exact matches only needed the start and end positions of the instance to make a complete tag, which was done automatically using regular expressions. Tags for the partial matches were done manually as the value in the original annotation did not match the instance in the note exactly.

Some annotations were straightforward, while others required additional work. For example, participants could choose the color ”yellow” for urine. However, since the relation to urine was conveyed in the subcategory, this relation was lost. To address this, additional entity types were created. For example the entity type ”OUT” (as something leaving the body), was created for words like ”urine” and ”stool”. The resulting tagset was designed to ensure that, if all entities were accurately identified and appropriately combined, the original structured annotation could be reconstructed. After settling on a tagset the process of tagging all notes began.

One person tagged the dataset, using approximately 20 hours. Every note was looked at four times. Beyond the notes that already had an annotations, every non-annotated note were tagged as well. A total of 23 entity types were used (Table 3).

5 Experiments

Extracting entities from the dataset could prove to be difficult. Some verbs, like ”walk”, belongs to different categories based on the tense of the word and the surrounding words. The word ”big” (”store” in Danish) is used both as a description of an AMOUNT *”The patient consumed two big portions of food”* or as a MODIFIER *”The patient have big problems eating”* (directly translated from Danish). Additionally, some entity types appear much less frequently than others, resulting in an unbalanced dataset where entities occur between 13 and 201 times. Lastly words like ”nasogastric tube” (nasalsonde in Danish) and ”Foley cathether” (KAD in Danish) are not common words and very specific to the medical domain, which might affect the results in a negative way.

5.1 Data split

Due to the size of the dataset, we used k-fold cross-validation for the evaluation. A value of k=6 was chosen, ensuring each entity type appears at

Tag	Description
PSYCHOLOGICAL	A psychological symptom (e.g. *sad, happy, angry, frustrated, confused*)
PHYSIOLOGICAL	A physiological symptom or condition (e.g. *constipated, nauseous, bound to bed*)
STATE	A state a patient can be in (e.g. *sleeping, sleepy, relaxed, awake*)
ASSISTIVE DEVICE	Items such as *walker, lift, hearing aids, diaper*
QUANTITY	A quantity defined numerically or textually (e.g. *4, 600, one, two*)
AMOUNT	A non-numerical amount (e.g. *big, small, large, huge, several*)
PERSONNEL	Any hospital personnel or outside personnel (e.g. *nurse, doctor, porter, ergotherapist, interpreter, he*)
PATIENT	Any mention of a patient (e.g. *Jack, William, pt, patient, him, her*)
IN	Anything that goes into a patient (e.g. *water, food, tubefood*)
OUT	Anything that goes out of a patient (e.g. *aspiration, stool, urine*)
CONSISTENCY	The consistency of OUT and IN (e.g. *soft, hard, liquid, gratin*)
UNIT	Units of measurement (e.g. *ml, mg, x*)
COMMUNICATION	Everything related to communication with the patient (e.g. *Danish, French, German, deaf, mute, reduced hearing*)
COLOR	Color of something (e.g. *brown, orange, red, green, yellow*)
APPEARANCE	The appearance of something (e.g. *clear, murky, dark*)
ACCESS	Access on the patient's body (e.g. *catheter, feeding tube, nasogastric tube*)
SOCIAL	Family members and friends (e.g. *daughter, son, neighbor, friend*)
MODIFIER	A word that modifies the meaning of a word (e.g. *much, less, very, good*)
NEGATION	A word that negates another word (e.g. *not, no*)
LOCATION	A location something can be (e.g. *bed, chair, toilet, leaf ear*)
TIME	An indication of time (e.g. *night shift, day shift, upon inspection, yesterday, tomorrow, after rounds*)
ACTION	An event that has happened (e.g. *eaten, mobilized, instructed, helped*)
ACTIVITY	An activity the patient can do or can be done to the patient (e.g. *walks, eats, drinks*)

Table 3: List of entities

least twice in every split. The data was stratified based on the entity tags for each note, maintaining roughly equal occurrences of entity tags and notes across splits.

5.2 Models

As the notes are in Danish, the number of models available for testing is limited.

5.2.1 BERTs

Four BERT models and one RoBERTa model will be tested.

- **bert-base-cased** (Devlin et al., 2019): An English BERT model not trained on Danish, tested here for comparison with Nordic language models.
- **danishBERT-uncased** (Certainly, 2023): A Danish BERT model trained on 9.5GB of text.
- **bert-base-swedish-cased** (KB (Kungliga Biblioteket), 2023): A Swedish BERT model trained on 15GB of text. Although Swedish, it has more training data than Danish models and it is cased.
- **nb-bert-base-cased** (Kummervold et al., 2021): A Norwegian model trained on the 48.9GB Norwegian Colossal Corpus, showing strong results for Danish tasks.
- **xlm-roberta-base-cased** (Conneau et al., 2019): A multilingual model based on RoBERTa, trained on 2.5TB of Common Crawl data, outperforming mBERT.

A token classification head was attached on top of the BERT/RoBERTa models, whereafter they were fine-tuned with the AdamW algorithm (Loshchilov and Hutter, 2019).

All models underwent a hyperparameter grid search optimization. The hyperparameters finetuned for included epoch $[15, 20, 25, 30, 35, 40, 45]$, learning rate $[2 \cdot 10^{-5}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}]$ and weight decay $[0.01, 0.1]$. Class weights were used in the loss function to handle the unbalanced classes.

5.2.2 Conditional Random Field

The Conditional Random Field (CRF) model developed for this study is supplied with a range of automatically computable features. These features include:

- Capitalization status of the current word, the preceding word, and the following word (uppercase and title case).
- Numeric status, identifying if the word consists of digits.
- Word2Vec embeddings from a Danish model (Sørensen, 2020), providing semantic representations for each word.

Additionally, the model identifies whether a word is at the beginning or end of a sentence, and it receives the same entity tags as the BERT models receive. The hyperparameters we optimized were c1 and c2 (the ℓ_1 and ℓ_2 regularization coefficients) $[0.01, 0.1, 0.5, 1.0]$ and the maximum number of iterations $[50, 75, 100]$.

5.3 Evaluation strategy

The BERT models and CRF model use the BIO (Beginning, inside, ouside) tag scheme and a prediction is only correct if the model predicts all B and I tags associated with an entity. A micro, macro and weighted avg f1 score is calculated for each model.

5.4 Results

Table 4 shows the average performance across all entities on the CRF model and the BERT models. The results for individual entity types and all tested models can be seen in Appendix A, Table 6. Not shown in any of the tables is the bert-base-cased model which achieved a macro f1 score of 0.613.

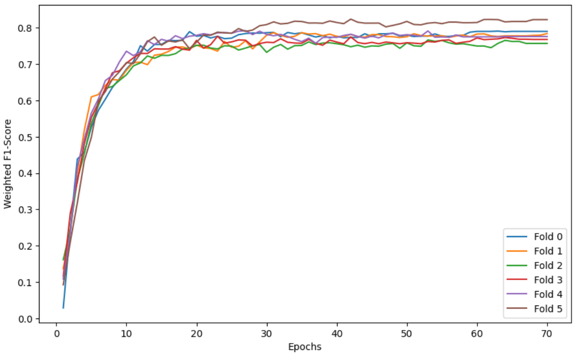


Figure 7: f1 for each epoch, with all 6 folds for DanishBERT

6 Discussion

6.1 Data collection and annotation

The note-writing aspect was successful, with most notes being of high quality and nuanced, indicating the web application's effectiveness. However, the annotation phase presented challenges, requiring significant effort to address low data quality, a common risk with crowdsourcing (Travis and Burton, 2023).

There are many reasons which could explain why the annotation part was less of an success and unfortunately the only feedback from the participants after the web application launched were a few comments on Facebook. Potential reasons for the troubles with the annotation part could be:

- The participants did not understand the task.
- The participants found the interface provided too difficult to use.
- The inherent problems in structured documentation (time consuming, hard to find the right categories) (Baumann et al., 2018).
- Too much to be expected from volunteers.

Our expectation was that the participants would quickly learn how to fill in the structured annotations, as the interface matched what is used in a real EHR, but the low quality of the annotations and notes without annotations suggested that this part remained difficult to use successfully.

There are several options to mitigate these issues:

- Improve the interface of the annotation process and put it through a more rigorous testing before beginning the data collection. This is time consuming, but could lead to better results.

	CRF	danishBERT	nb-BERT	xlm-roBERTa-base	swedishBERT
micro avg	0.740 ± 0.033	**0.779 ± 0.018**	0.750 ± 0.033	0.763 ± 0.037	0.725 ± 0.029
macro avg	0.704 ± 0.044	**0.744 ± 0.018**	0.739 ± 0.042	0.732 ± 0.030	0.699 ± 0.031
weighted avg	0.726 ± 0.038	**0.783 ± 0.016**	0.771 ± 0.032	0.772 ± 0.031	0.736 ± 0.029

Table 4: A comparison between CRF and the BERT models, with average f1 score over a 6-fold-cross validation run and standard deviation between those runs. The best results are bolded.

- Pay nurses and give more detailed instructions. This is expensive, but would provide better quality as the annotators are better instructed.
- Lastly the annotation part of the process could be removed, leaving only the write note part, which could lead to more notes as it is an easier task and thus more encouraging for the participants. However, doing this would lead to more work, as some of the annotations done by the participants were directly usable.

The dataset does not cover all nurse-relevant problem areas, and even the represented nurse-relevant problem areas are incomplete. This limitation poses a challenge in evaluating the results, as there might be nuances of nurse documentation that is harder to capture than others.

Furthermore, the decision to discard annotations based on interpretation in favor of framing the task as a NER task, inadvertently contributes to the incompleteness in capturing the full spectrum of nursing documentation.

6.2 Information extraction

This section will discuss the results in regards to extracting entities from the dataset. When observing the results, one should take into consideration the high variance in the f1 scores between folds. Some folds, as illustrated in Figure 7 had a big difference in f1 score, which both highlights the importance of using a cross-validation strategy, but also indicates that the results might look different if the dataset were larger and more balanced. When looking at the results of this study, these things should be kept in mind.

The best model was the DanishBERT achieving a macro f1 of 0.744. As expected the nb-BERT, which has been shown to have solid performance on danish , showed similar performance with a macro f1 of 0.739 and achieved best performance on 8 entities, compared to the danish which had the best score on only 4 entities. The xlm-roBERTa-base (multilingual) had a solid performance as well with a macro f1 of 0.732 and best performance on 6 entities. SwedishBERT only managed a macro f1 of 0.699.

The CRF model performed well and performed best of all models in 7 entity types and only having a slightly lower macro f1 of 0.704. However, it did fall short completely on more entities than the BERT models, indicating that the more computational BERT models are more robust in their performance.

7 Conclusion

This study aimed to bridge the gap between structured and free-text documentation in healthcare using NLP techniques. The initial step involved constructing a dataset, which was necessary due to the absence of pre-existing suitable datasets in this domain. Following dataset construction, the study focused on extracting relevant information from nursing documentation within this newly created dataset.

The creation of a synthetic dataset of annotated nurse notes was accomplished through a web application. This application presented various stimuli to participants, prompting them to write corresponding notes. Subsequently, participants annotated their notes using categories reflective of those used in actual EHRs. Overall, the quality of the notes was high, although not all annotations were usable. A manual process was employed to eliminate incorrect annotations and convert the annotations into pairs of (entity type, entity). Additional support entities were manually added, ensuring that every word relevant to nurse documentation was properly tagged.

The process of extracting meaningful information from nurse documentation was approached as a NER task. Performance evaluation revealed that the the Danish, Norwegian and multilingual models had similar performances, with the best being the Danish which achieved a macro f1 score of 0.744, surpassing the CRF model, which scored 0.704. This performance difference highlights the necessity and efficiency of more advanced models like BERT in handling complex NER tasks.

However, it is important to note that the entity type/entity instance pairs extracted through this NER process do not directly correspond to the structured format which is used in EHRs. This gap underscores a potential area for future research, where the focus could be on transforming these pairs into EHR-compatible triples. Such a transformation is crucial for the practical application of this research in real-world EHR systems, potentially facilitating smoother integration of automated NLP-based documentation tools into healthcare workflows. Nevertheless, this study demonstrates that it is possible to generate synthetic nurse notes and extracting information relevant to nurse documentation from them.

8 Ethical Considerations

Our approach mitigates privacy concerns by using fictitious data, thereby reducing the risk associated with real patient information. However, there is a potential concern regarding the applicability of findings derived from this synthetic dataset, as the data may not accurately reflect real-world.

9 Limitations

With only 98 nurses participating in the study, the dataset is relatively small and only encompass a subset of possible nurse-related categories, potentially limiting its representativeness. Additionally, the lack of multiple reviewers for note quality assessment and the absence of inter-annotator agreement values for the entities diminish the robustness of the results. Lastly it is important to note that all of the participants' status as nurses cannot be verified, as the Facebook group used does not authenticate group members credentials.

References

Edward P. Ambinder. 2005. Electronic health records. *J Oncol Pract*, 1(2):57–63.

Lisa Ann Baumann, Jannah Baker, and Adam G. Elshaug. 2018. The impact of electronic health record systems on clinical documentation times: A systematic review. *Health Policy*, 122(8):827–836.

Maj-Britt Brinkmann, Birgitte Brask Skovgaard, and Raymond Kolbæk. 2020. Dokumentation er en væsentlig del af sygeplejen. *Fag & Forskning*, nr. 1:64–69.

Certainly. 2023. Certainly has trained the most advanced danish bert model to date. https://certainly.io/blog/danish-bert-model/. Accessed: 2023-12-12.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, Lu Shen, and et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Ralph Johnson. 2016. A comprehensive review of an electronic health record system soon to assume market ascendancy: Epic®. *Journal of Healthcare Communications*, 01.

KB (Kungliga Biblioteket). 2023. Bert-base, swedish, cased. https://huggingface.co/KB/bert-base-swedish-cased.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a Norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

MY Landolsi, L Hlaoua, and L Ben Romdhane. 2023. Information extraction from electronic medical documents: State of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516. Epub 2022 Nov 8.

S. J. Lewis and K. W. Heaton. 1997. Stool form scale as a useful guide to intestinal transit time. *Scandinavian Journal of Gastroenterology*, 32(9):920–924. PMID: 9299672.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv:1711.05101 [cs.LG]*. Published as a conference paper at ICLR 2019.

S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Styrelsen for Patientsikkerhed (SFPS). 2023. Sygeplejefaglig journalføring.

Nicolai Hartvig Sørensen. 2020. Word2vec-model for danish. https://korpus.dsl.dk/resources/details/word2vec.html. Accessed: [2023-12-12].

Emma Tram. 2017. Sygeplejerskers dokumentationspraksis. *Sygeplejersken*, 2017(11):26–27.

Jack Travis and Scot Burton. 2023. Do you trust what the survey says? examining data quality on online crowdsourcing platforms. *Walton College Insights*.

A Tables

	Models				
	XLM-BERT	DanishBERT	swedishBERT	nb-BERT	CRF
dropout	0.1	0.1	0.1	0.1	-
architecture	RoBERTaForTokenClassification	BertForTokenClassification			-
embedding	RoBERTa$_{\text{base}}$	BERT$_{\text{base}}$	BERT$_{\text{base}}$	BERT$_{\text{base}}$	-
parameters					
epoch	35	35	45	45	-
learning_rate	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	-
batch_size	8	8	8	8	-
weight_decay	0.1	0.01	0.01	0.1	-
c1	-	-	-	-	0.01
c2	-	-	-	-	0.01
max_iter	-	-	-	-	50
algorithm	-	-	-	-	lbfgs

Table 5: Training parameters

	Models				
Entity type	CRF	DanishBERT	nb-BERT	xlm-roBERTa	average support
PATIENT	0.948 ± 0.028	**0.963 ± 0.035**	0.942 ± 0.021	0.926 ± 0.072	33.5
PSYCHOLOGICAL	0.579 ± 0.034	0.782 ± 0.063	**0.805 ± 0.058**	0.789 ± 0.098	17.0
ASSISTIVE DEVICE	0.777 ± 0.064	0.814 ± 0.050	0.821 ± 0.051	**0.836 ± 0.081**	16.8
QUANTITY	0.921 ± 0.104	0.908 ± 0.022	0.931 ± 0.059	**0.955 ± 0.031**	16.8
ACTION	**0.764 ± 0.058**	0.713 ± 0.081	0.675 ± 0.084	0.642 ± 0.155	14.5
PHYSIOLOGICAL	0.368 ± 0.151	0.551 ± 0.045	**0.615 ± 0.080**	0.511 ± 0.135	14.2
TIME	0.502 ± 0.129	0.604 ± 0.147	**0.608 ± 0.101**	0.602 ± 0.085	11.3
UNIT	**0.968 ± 0.044**	0.941 ± 0.054	0.885 ± 0.109	0.926 ± 0.085	11.2
OUT	0.712 ± 0.084	0.804 ± 0.041	0.863 ± 0.085	**0.869 ± 0.079**	11.0
MODIFIER	0.510 ± 0.214	**0.688 ± 0.115**	0.592 ± 0.132	0.580 ± 0.148	9.8
ACTIVITY	0.661 ± 0.127	0.650 ± 0.125	0.643 ± 0.099	**0.714 ± 0.094**	9.0
STATE	0.822 ± 0.111	0.903 ± 0.108	**0.908 ± 0.081**	0.904 ± 0.132	7.5
PERSONNEL	0.761 ± 0.180	0.774 ± 0.178	**0.852 ± 0.116**	0.794 ± 0.138	7.2
IN	0.635 ± 0.184	**0.766 ± 0.109**	0.013 ± 0.030	0.630 ± 0.125	6.7
AMOUNT	0.617 ± 0.114	0.702 ± 0.116	0.761 ± 0.115	**0.765 ± 0.095**	6.3
CONSISTENCY	0.713 ± 0.111	0.707 ± 0.143	0.721 ± 0.101	**0.770 ± 0.111**	5.5
COMMUNICATION	0.513 ± 0.287	0.588 ± 0.289	**0.760 ± 0.181**	0.643 ± 0.312	4.8
ASSIS/LOCATION	0.745 ± 0.203	0.843 ± 0.033	**0.850 ± 0.150**	0.736 ± 0.187	4.0
ACCESS	**0.825 ± 0.108**	0.806 ± 0.196	0.708 ± 0.220	0.747 ± 0.221	3.5
COLOR	**0.900 ± 0.200**	0.856 ± 0.245	0.883 ± 0.186	0.867 ± 0.221	3.3
SOCIAL	**0.960 ± 0.080**	0.867 ± 0.094	0.952 ± 0.067	0.875 ± 0.191	3.2
LOCATION	0.280 ± 0.232	0.000 ± 0.000	**0.436 ± 0.261**	0.000 ± 0.000	2.8
NEGATION	**0.867 ± 0.163**	0.778 ± 0.050	0.704 ± 0.137	0.721 ± 0.134	2.8
APPEARANCE	0.560 ± 0.285	**0.856 ± 0.151**	0.800 ± 0.224	0.759 ± 0.214	2.2
micro avg	0.740 ± 0.033	**0.779 ± 0.018**	0.750 ± 0.033	0.763 ± 0.037	222.333
macro avg	0.704 ± 0.044	**0.744 ± 0.018**	0.739 ± 0.042	0.732 ± 0.030	222.333
weighted avg	0.726 ± 0.038	**0.783 ± 0.016**	0.771 ± 0.032	0.772 ± 0.031	222.333

Table 6: A comparison between different models, with average f1 score over a 6-fold-cross validation run and standard deviation between those runs. The best result being bolded. swedishBERT not shown.

Category	SubCategory	SubSubCategory
Functional Level	Current functional level	Mobility aids: 37
		Mobility assistance: 18
		Assistance with elimination: 3
		Mobility restrictions: 3
		Personal hygiene assistance: 2
	Habitual functional level	Habitual mobility: 3
		Mobility aids: 2
		Personal hygiene assistance: 2
	Mobilization activity	Mobility aids: 30
		Mobility assistance: 17
		Mobilization (number of times): 6
		Mobilization (where the patient is mobilized to): 6
		Mobilization (distance) in meters: 3
		Mobilization (time): 1
Sleep and rest	Habitual sleep	Sleep pattern: 2
		Sleep disturbances: 6
	Rest	Resting state: 9
	Sleep registration	Hours slept during shift: 18
		Sleep quality: 8
		Current state: 8
	Sleep/Rest issues	Problems: 23 Measures taken: 7

Table 7: Annotations 1/3

Category	SubCategory	SubSubCategory
Communication	Barriers	Language: 12
		Hearing: 10
		Cognitive: 8
	Communication assistance	Technical aids: 22
		Need for interpreter: 17
		Need for relatives: 9
Psychological and social	Psychological	Current mental state: 55
		Reaction to illness: 11
		Habitual mental state: 4
		Illness insight: 4
		Perception of health: 1
	Social	Network: 8
Elimination	Aspiration	Amount: 7
		Frequency: 5
		Color: 3
	Stool registration	Consistency: 16
		Amount: 15
		Frequency: 14
		Color: 12
		Location: 2
	Stool status registration	Stool status: 8
	Urination registration	Amount in ml: 13
		Source: 12
		Appearance: 11
		Color: 10
		Amount: 9
	Regular bowel movements	Frequency: 4 Consistency: 3

Table 8: Annotations 2/3

Category	SubCategory	SubSubCategory
Nutrition	Current nutritional status	Weight in kg: 2
		Height in cm: 1
	Assistance to eat and drink	Assistance to drink: 3
		Assistance to eat: 2
	Diet	Consistency food: 10
		Diet: 5
		Consistency liquids: 4
	Issues	Nausea: 11
		Appetite: 9
		Swallowing difficulties: 1
	Meal registration	Percentage of intake: 13
		Problems: 5
		Intake via tube as planned: 5
		Intake via tube in ml: 3

Table 9: Annotations 3/3

B Description of stimuli

1. A 20-second video of a man trying to eat food in a kitchen, but ends up pushing it away while frowning.
2. A 20-second video of a man enjoying a sandwhich outside.
3. A picture of an elderly woman receiving food through a nasogastric feeding tube.
4. A picture of an elderly woman walking with a walker in a park.
5. A picture of two healthcare professionals using a ceiling hoist to mobilize a man in a hospital bed, with a wheelchair at the end of the bed.
6. A 15-second video of a 100-year old woman running.
7. A picture of a healthcare professional assisting a man using a walker.
8. A picture of two healthcare professionals assisting a man walking with elbow sticks.
9. A picture of a man placing a hearing aid in an ear.
10. A video of a young woman using sign language.
11. A video of an interpreter translating Spanish in a hospital setting.
12. A picture of a man lying in a hospital bed, with another man in non-uniform clothing and a doctor standing besides it.
13. A picture of a happy smiling woman in a hospital gown in a bed.
14. A picture divided in two: To the left a doctor speaking and gesturing with his hands, to the right a man putting his hands pressed against his head and his face and his brow deeply furrowed.

15. A picture divided in two: To the left a doctor speaking and gesturing with his hands, to the right a man with visible tears on his face.

16. A picture of a woman lying in a bed with eyes closed in a dimly lit room.

17. A drawing of a man lying in bed counting sheeps.

18. A 10-second video of a young man walking around restlessly.

19. A drawing of bacteria, with the names of three bacteria known to cause diarrhea.

20. A picture of a diaper.

21. A picture of a person on a toilet.

22. A picture of a urine drainage bag.

23. A 10-second video clip of a woman vomiting in a bag in a restaurant.