Journal of the American Medical Informatics Association, 27(1), 2020, 81–88
doi: 10.1093/jamia/ocz150
Advance Access Publication Date: 12 October 2019
Research and Applications

Research and Applications

Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods

Hans Moen, [1,†] Kai Hakala, [1,†] Laura-Maria Peltonen, [2] Henry Suhonen, [2,3] Filip Ginter, [1] Tapio Salakoski, [1] and Sanna Salanterä [2,3]

[1]Department of Future Technologies, University of Turku, Turku, Finland, [2]Department of Nursing Science, University of Turku, Turku, Finland, and [3]Department of Nursing, Turku University Hospital, Turku, Finland

[†]The first two authors contributed equally .
Corresponding Author: Hans Moen, PhD, Department of Future Technologies, University of Turku, Vesilinnantie 5, 20500 Turku, Finland; hans.moen@utu.fi

Received 21 February 2019; Revised 4 July 2019; Editorial Decision 19 July 2019; Accepted 3 August 2019

ABSTRACT

Objective: This study focuses on the task of automatically assigning standardized (topical) subject headings to free-text sentences in clinical nursing notes. The underlying motivation is to support nurses when they document patient care by developing a computer system that can assist in incorporating suitable subject headings that reflect the documented topics. Central in this study is performance evaluation of several text classification methods to assess the feasibility of developing such a system.
Materials and Methods: Seven text classification methods are evaluated using a corpus of approximately 0.5 million nursing notes (5.5 million sentences) with 676 unique headings extracted from a Finnish university hospital. Several of these methods are based on artificial neural networks. Evaluation is first done in an automatic manner for all methods, then a manual error analysis is done on a sample.
Results: We find that a method based on a bidirectional long short-term memory network performs best with an average recall of 0.5435 when allowed to suggest 1 subject heading per sentence and 0.8954 when allowed to suggest 10 subject headings per sentence. However, other methods achieve comparable results. The manual analysis indicates that the predictions are better than what the automatic evaluation suggests.
Conclusions: The results indicate that several of the tested methods perform well in suggesting the most appropriate subject headings on sentence level. Thus, we find it feasible to develop a text classification system that can support the use of standardized terminologies and save nurses time and effort on care documentation.

Key words: natural language processing, electronic health records, nursing documentation, text classification, clinical decision support

INTRODUCTION

The documentation of care in hospitals is important for supporting a safe care continuity. This includes documenting information about patients' health, administered care, and future care plans. Performing this documentation constitutes a relatively large portion of the work conducted by clinicians, leaving less time for direct patient care. According to literature, nurses spend up to 35%, with an average of 19%, of their working time on documentation.[1] In addition, the number of items being documented is increasing steadily.[2] In many countries, nurses are required to conduct some type of structuring of the patient information they enter into electronic health record (EHR) systems.[3] Such structuring methods include the use of documentation standards, classifications, and standardized terminologies.[4] In the hospital district in Finland from where the

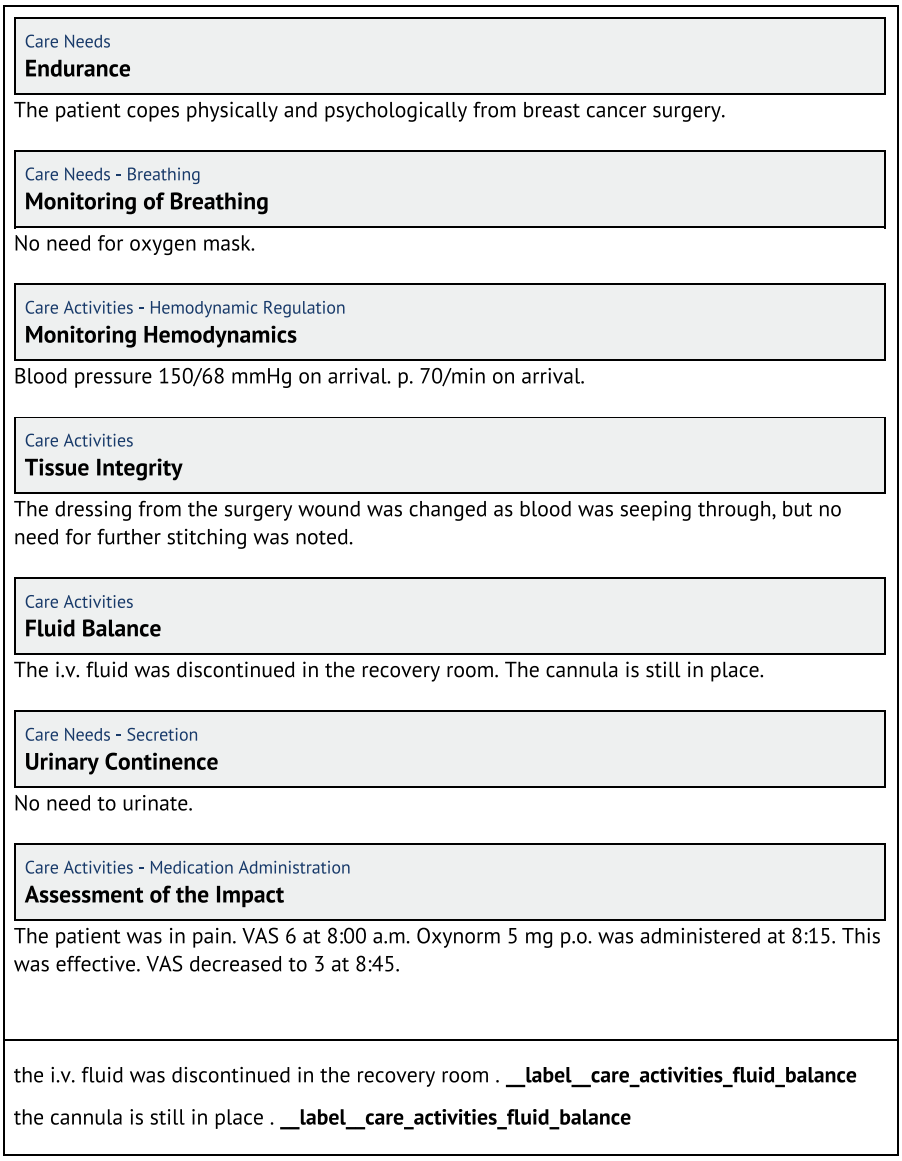

Figure 1. The upper part shows an example of a nursing note. The lower part shows how a paragraph is split into individual training examples, each consisting of a sentence with an attached subject heading label. In this way we turn all sentences, from all paragraphs, in the dataset into training examples. Translated to English from Finnish.

data used in our experiment originate (see Data section), nurses are required to plan and divide the information they want to document into paragraphs and label them with standardized subject headings (also referred to as topical subject headings, or simply headings) (see the upper part of Figure 1 for an example of a nursing note). With potentially a large set of subject headings to choose from, selecting the correct headings can be challenging, as it requires nurses to learn a vast hierarchy of terms by heart.[5] As an example, our dataset contains 676 unique headings. This contributes to making the documentation process more time consuming compared with using fully unstructured free (narrative) text. Further, correct use of such subject headings requires training.[5,6]

Natural language processing methods have the potential to support clinicians in structuring the information they document.[7] In the presented experiment, we approach the task of having the computer automatically suggest subject headings for the text in nursing notes. We want the computer to automatically classify the text on the level of sentences by selecting or suggesting subject headings that best describe the information that the sentences represent. For example, given the sentence "Taken care of the cannula himself," we would ideally like to have the computer suggest a heading like "Cannula Care," which is one of the subject headings found in the underlying documentation standard. To assess whether or not this is feasible to do, and what classification method is best suited for this task, we apply and compare 7 different text classification methods—primarily supervised machine learning methods.

The main motivation for our work is to develop a system that can assist nurses in using subject headings when they document. We see 2 use cases for such a system: (1) it helps by suggesting the most appropriate headings at the time of documentation (eg, for each sentence or paragraph) and (2) it retrospectively structures and assigns subject headings to nursing (shift) notes that are written as

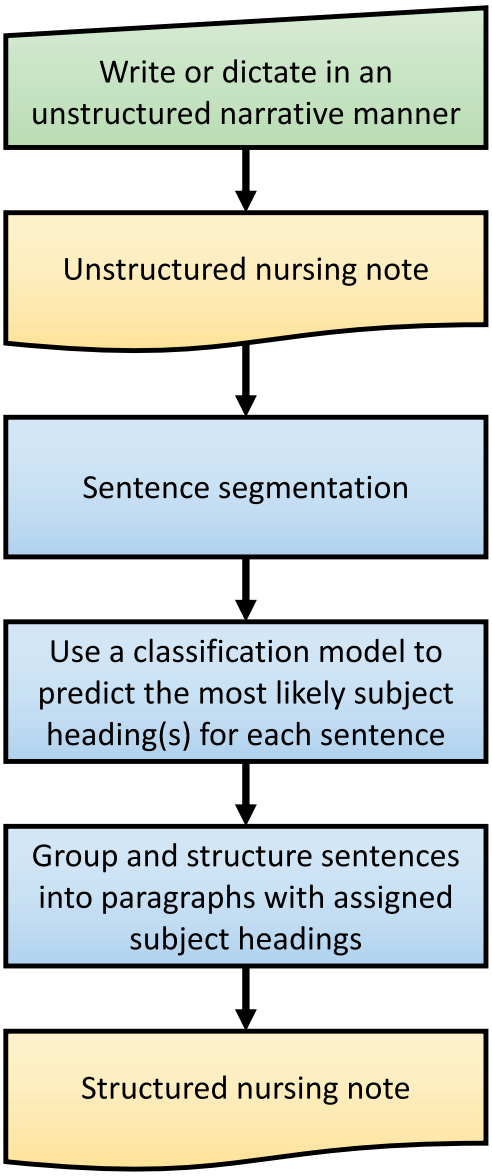

Figure 2. Flowchart showing the intended workflow in which a classification model is used to assign subject headings to sentences before restructuring them into paragraphs. The main focus of this study is to identify the best-performing method or model for the classification task.

free (narrative) text without having planned and structured the text with respect to sections and standardized subject headings. In the second use case, such a system would allow nurses to primarily focus on writing or dictating the full story. This would ensure that "weaker signals" and more vague concepts and expressions are documented, such as decision-making processes, uncertainty, and probabilities, which tend to be easier to document using free (narrative) text compared with more structured representations.[7] Further, in this second use case, a structured version will be automatically generated and available alongside the original text. Figure 2 illustrates how we see the workflow in this second use case. We hypothesize that such a system will allow nurses to save time and effort in tasks related to care documentation. Consequently, this will free up more time to concentrate on the patient and deliver better care.[8] Another possible outcome of using such a system is improved consistency and correctness in nurses' use of subject headings.

Text classification, or text class prediction, concerns the task of assigning 1 or more predefined classes to text of various length. We have only identified a few papers focusing on supporting the assignment of standardized headings to free-text sections in clinical notes.[9–11] In terms of classifiers these rely on naive Bayesian classifiers,[9] hidden Markov models,[10] and Bayesian networks.[11] However, they all use <30 unique headings, and the input text is assumed to be a linear sequence of sections. Another closely related and well-studied research topic is the assignment of diagnosis codes, such as the International Classification of Diseases, to the target documents and care episodes, but these studies tend to focus on document level classification.[12,13] Text classification for clinical text has also been used in other tasks such as mapping of phrases to medical concepts,[14] classifying patients smoking status,[15] classifying obesity and its comorbidities,[16] and classifying heart disease risks.[17] The latter 3 tasks have been shared tasks arranged by the Informatics for Integrating Biology and the Bedside (i2b2) center.

Since deep learning with artificial neural networks has recently achieved state-of-the-art performance in text classification in other domains, see Zhang et al[18] and Tang et al,[19] we now want to study how they perform in this domain for the described task and data. Further, given the unique features of clinical text such as unfinished sentences, missing subject, abbreviations and mixture of Latin, spoken language, and organization-specific terminology, performance evaluations of such methods tested on text and comparable tasks in other domains are not necessarily representative.

MATERIALS AND METHODS

Data

The dataset consists of de-identified clinical nursing shift notes extracted from a Finnish university hospital EHR system. Selection criteria was patients with any type of heart-related problem in the period of 2005-2009. Nursing notes from all units visited during their hospital stay are included. Ethical approval was obtained from the hospital district's ethics committee (17.2.2009 §67) and research approval was obtained from the medical director of the hospital district (2/2009).

The dataset spans across the transition phase of 2 nursing documentation standards, the latter being the latest version of the Finnish care classification standard (FinCC)[5] and the former being an earlier iteration of a similar standard. This means that the dataset contains a mix of subject headings from both standards. FinCC consists primarily of the Finnish classification of nursing diagnoses/care needs (FiCND) and the Finnish classification of nursing interventions/care activities (FiCNI) and is based on the international Clinical Care Classification System. Both of these have 3 taxonomic levels, with 545 subject headings altogether.

The total number of nursing notes in the dataset is approximately 0.5 million. In this study, we only include paragraphs having a subject heading. Further, we decided to exclude headings occurring <100 times, resulting in a total of 676 unique subject headings. Their frequency count ranges from 100 to 222 984, with an average of 4896. See upper part of Table 1 (count ≥100) for an insight into the most and least common subject headings. The excluded headings, occurring <100 times, constitute only about 1% of the total paragraphs, and 1% of the sentences (ie, training examples). One reason for this cutoff is to ensure that we have a fair amount of training data for each subject heading. A second reason is that many of the subject headings occurring <100 times are mostly normal sen-

Table 1. The 6 most and least common headings in the dataset, and 10 headings that occur <100 times, meaning that they were excluded.

Subject heading	n	%
Wellness and ability to function	222 984	6.737
Physiological measurements	198 919	6.010
Nutrition	135 984	4.109
Urinary tracts	128 486	3.882
Activity	123 294	3.725
Medication	117 502	3.550
Urinary incontinence	101	0.003
Change in the kidney and urinary tract activity	101	0.003
Other	101	0.003
Neuropathic pain	100	0.003
Coping with activities of daily living	100	0.003
Loss of appetite	100	0.003
Chemtherapy (not chemotherapy)	99	—
Excretion	99	—
Intravenous alimentation	98	—
Oral and mucus related patient education	98	—
Organization of sequel physiotherapy	98	—
Urination disorders	97	—
Lung function associated with breathing	97	—
Taking a sample	97	—
Supporting communication	96	—
Level of Consciousness Glascow (not Glasgow) Coma Scale	96	—

Translated to English from Finnish.

tences that have been erroneously written into the heading field or customized versions of the more common ones, with a tendency to contain spelling errors (see lower part of Table 1). Further, it is worth noting that the latest version of FinCC only contains 545 unique headings.

There are approximately 5.5 million sentences in the dataset, in which the average sentence length is 7 tokens (tokens are here defined as the space separated units in the sentence), with an average of 2.1 sentences per paragraph. It contains 133 890 unique tokens and approximately 38.5 million tokens in total. Figure 1 (top) shows an example of a nursing note. We have split the paragraphs into individual training examples to enable sentence-level multiclass classification, as illustrated in Figure 1 (bottom). The training examples (sentence + subject heading) were split into training, development. and test sets with 60%, 20%, and 20% of the data using stratified sampling, respectively.

Methods

The following methods and baselines were used as many of them have been shown to perform well in text classification and used in several recent studies.[18–21] Model architectures and hyperparameters for the below methods can be found in the Supplementary Appendix.

`LSTM` and `BidirLSTM`

Long short-term memory (LSTM) networks are a form of recurrent neural network (RNN) architecture which are able to process sequential data in which each decision is influenced by the previous observations.[22,23] In our case, one sequence is a sentence, given to the network one word at a time. In the case of bidirectional LSTMs, the network reads the input sentence from both directions. Internally LSTMs maintain a state the network can modify by forgetting parts of the old information and adding something new from the current input. Thus, while reading the input words, an LSTM network has the ability to store semantic information of each relevant word and will make the final classification decision based on a semantic representation of the whole sentence.

`CNN`

Convolutional neural networks (CNNs)[24] consist of a set of convolutional kernels, each kernel trying to detect a small but relevant pattern from the given input data. These kernels are then applied to the data in a sliding window manner. In our case, they are applied over the sequence of words in a sentence, with the window limiting the scope of the kernel to only a few nearby words. For each position in the input sentence a kernel measures how well the currently observed window (ie, a short word sequence) fits to the pattern that the kernel is trained to detect. Thus, in our case the goal for each convolutional kernel is to detect short phrases within the sentence that might be relevant for the topical classification. The benefit of a convolutional network is that the classification decisions are based on small fractions of the sentence, possibly occurring within various sentence structures. This differs from RNNs that try to capture the semantics of the whole sentence, which in practice can be hard to achieve.

`fastText`

The FastText library is a neural network-based text representation[25] and classification package.[20] Compared with the `LSTM` and `CNN` methods, fastText represents a simpler and faster approach to text classification. Still, it has been reported to achieve state of the art results.[20] FastText utilizes a simple feedforward neural network with input layer reflecting the vocabulary in the corpus and the input sentence is represented as a bag of words (BoW) occurring in the sentence. To include some word order information, the input layer can also be configured to take word n-grams (in addition to unigrams).

`BoWLinearSVC` (**baseline**)

As a baseline approach we use a linear support vector machine classifier (SVM)[26,27] with TF-IDF (term-frequency inverse-document-frequency)[28] weighted BoW representation of the dataset as feature vectors. We also experimented with using word bigrams and trigrams in addition to the BoW features, but did not observe any performance increase.

Whereas SVMs are linear in nature, they can be used to produce nonlinear hypotheses with kernels, most commonly with the radial basis function (RBF) kernels.[27] As the RBF SVMs are computationally demanding, training such models with the magnitude of data we are dealing with is not possible. However, we have tested training an RBF SVM model for the task using kernel approximations.[29,30] Unfortunately, we were not able to produce a computationally feasible approximation that indicated a strong performance and thus this approach has been omitted from the detailed evaluation.

`RandomForest` (**baseline**)

Another baseline that we test is random forest classifier, which has shown generally good performance on a variety of classification tasks.[31] The same BoW features are used as with `BoWLinearSVC`.

`Word&HeadingEmbeddings` **(baseline)**
We train sentence and heading vectors or embeddings using the word2vec neural network distributional semantics tool/library.[32] The aim of the word2vec tool is to represent each word seen in a large text corpus with a high-dimensional vector representation encapsulating the semantic and syntactic features of the word. These vectors are learned by analyzing the contexts in which these words commonly appear.[33] For classification we first compose sentence and heading vectors by simply summing the trained word vectors (first normalized to unit length), which has been shown to preserve the semantic characteristics of the sentences,[32] and has been used in medical term normalization.[34] Then, for each sentence, the most similar headings are selected through calculating cosine similarity between sentence vectors and heading vectors. Owing to the unsupervised nature of this method, its main advantage over the supervised approaches is that new subject headings can be added without having to retrain the classification pipeline. These same word embeddings are utilized in the `CNN`, `LSTM` and `BidirLSTM` models.

Naive baselines: `MostCommon` **and** `Random`
We also use 2 simple baselines: one that always suggests the most frequent headings, `MostCommon`, and another that selects headings randomly from a uniform distribution, `Random`.

RESULTS

The selected methods were trained on the training data described in the Data section, with hyperparameters selected to maximize the performance on the development set. The test set was used when generating the reported results.

Automatic evaluation

Automatic evaluation was conducted by simply checking, for each sentence, whether or not the methods were able to suggest the same subject headings as those assigned by the nurses in the original text. We will refer to the originally assigned headings as *gold headings*. We believe that a system designed to assist nurses in selecting headings should suggest only a couple of headings to actually simplify the selection task, and in this evaluation we use 10 as the maximum. Results were calculated per method as average recall at top N (R@N) and as mean reciprocal rank (MRR) for the top 10 suggested headings per sentence. R@N for a method is calculated by first having it suggest a set of N headings for each sentence in the test set. For each sentence, if the gold heading is in the corresponding set of suggestions, it is assessed as a correct classification. As an example, take the sentence, "The surgery wound looks good." Now we have one method predict 10 headings. Among these, let us say that the correct/gold heading "Tissue Integrity" is at position 3. Because the correct heading is among the top 10 predicted headings, we consider the classification correct in the R@10 setting. The final R@10 score for the method is the sum of the correctly classified sentences divided by the number of sentences in the dataset. Recall was calculated individually for N values in the range 1-10 (ie, R@1, R@2, . . ., R@10, where R@1 is equal to accuracy). MRR is the average multiplicative inverse of the rank of the correct heading in the top 10 list retrieved for each sentence. MRR is sensitive to ranking, so methods that rank the correct headings first receive higher MRR scores. Given the example sentence above, its MRR score is calculated as 1/3, where 3 is the position of the gold heading in the list of predicted headings. For each method we report the average score over all sentences in the dataset.

Table 2. R@1 (ie, accuracy score) and R@10, as well as the MRR for each method.

Method	R@1, Accuracy	R@10	MRR
`BidirLSTM`	0.5435	0.8954	0.6621
`LSTM`	0.5429	0.8932	0.6610
`CNN`	0.5348	0.8856	0.6526
`fastText`	0.5224	0.8801	0.6428
`BoWLinearSVC`	0.5149	0.8486	0.6286
`RandomForest`	0.4896	0.7690	0.5868
`Word&HeadingEmbeddings`	0.1629	0.5111	0.2633
`MostCommon`	0.1038	0.3776	0.1679
`Random`	0.0015	0.0150	0.0044

MRR: mean reciprocal rank; R@1: recall at 1 subject heading per sentence; R@10: recall at 10 subject headings per sentence.

Table 2 shows R@1, R@10, and MRR scores for each method. `BidirLSTM` performed best according to these measures. However, `LSTM` achieved close to equal scores and all neural models performed better than more traditional `BoWLinearSVC` and `RandomForest` approaches. All performance differences between methods were found to be statistically significant ($P < .05$). The MRR score for `BidirLSTM` indicates that the correct heading is on average found between the first and second suggestion (mean rank = 1.51). However, the other top 4 methods show very similar ranking properties.

Figure 3 shows R@1-10 for each method plotted in a graph. R@N scores in the interval R@1 to R@10 for each method are following similar curves, this is particularly the case for the top 5 performing methods.

Manual error analysis

Based on information provided by domain experts, we hypothesize that the nurses might not always be using the correct subject headings when they document. As they often have limited time to do the documentation, the headings they use might not always correctly describe the information written underneath. This is likely to have an impact on the training of the methods or models and negatively affect the automatic evaluation scores. We further hypothesize that quite a few of the sentences that were incorrectly classified according to the automatic evaluation could be correct according to human domain experts. In addition to the quality of the training data, one obvious error source is that the data stem from 2 different documentation standards. This could result in the predicted subject headings being very similar to, but not exactly the same as, the gold headings. There could also be cases where the classifier has predicted subject headings from a different level in the FinCC hierarchy compared with the gold heading.

To test these hypotheses and to get a better understanding of the actual performance of the system, we decided to manually evaluate a sample of the data. We picked 200 randomly selected sentences and evaluated these with respect to (1) their headings originally assigned by the nurses and (2) their top 1 predicted headings by `BidirLSTM`. These were both evaluated separately by 2 domain experts, while a third was involved in deciding the consensus where they disagreed. The evaluators assigned each sentence-heading pair (original and predicted) to 1 of the following 4 classes:

1. **Correct** - the subject heading is correct for this sentence.

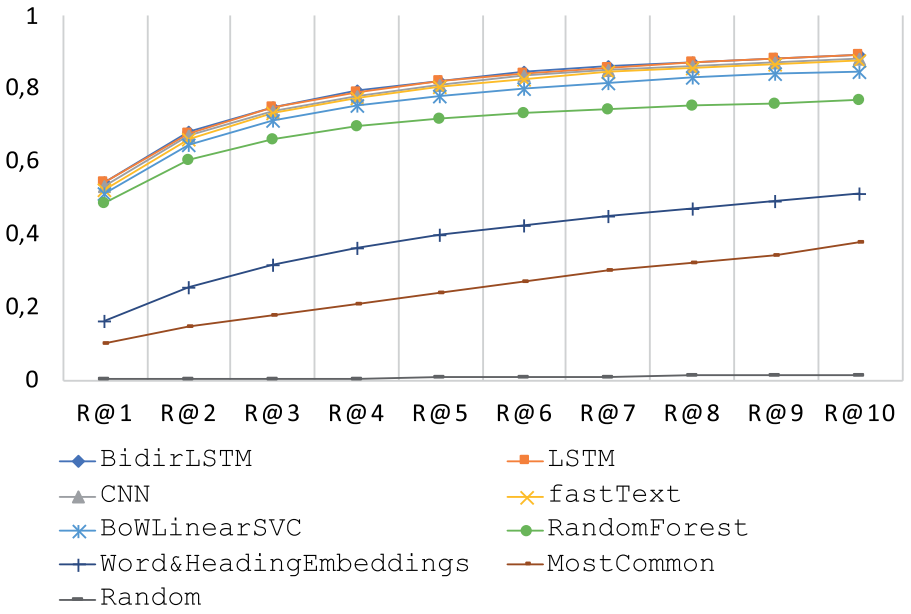

Figure 3. Recall at (R@) 1-10 plotted for each method. A high quality version of the plot for the top 6 performing methods is included in the Supplementary Appendix.

2. **Maybe correct** - the subject heading could be correct, but I am unable to say for sure without seeing the rest of the nursing note.
3. **Incorrect** - the subject heading does not seem to correspond well with the given sentence.
4. **Unable to assess** - unable to assess whether or not this subject heading is correct.

The result from this analysis can be seen in Table 3. The analysis indicates that between 74% and 89% (1+2) of the originally assigned headings are correct. In other words, between 11% and 26% of the gold headings are not correct. It also shows that between 81.50% and 94.50% of the headings predicted by `BidirLSTM` are correct according to the evaluators. In comparison, this amounts to only 58.00% correct when applying the R@1 automated evaluation metric (accuracy [ie, the percentage of sentences whose predicted heading equals the heading originally assigned by a nurse]). Further, the latter score is comparable to the automatic evaluation on the full test set, where this method predicts the correct heading for 54.35% of the sentences (Table 2).

DISCUSSION

The results show promising performance for many of the tested methods, especially so when they are allowed to suggest 10 subject headings. The best-performing method is `BidirLSTM`. We observe that the performance differences—the order of performance in particular—seems to be comparable to what is reported in.[20] The MRR scores indicate that the order in which these methods rank the correct headings does not differ much among the top-performing methods. This experiment also shows that the simpler and faster to train `fastText` method can be used without much loss in performance compared with `(Bidir)LSTM` and `CNN`. It is worth noting that CNN-based models have commonly surpassed RNNs and attentive models in short text classification outside the clinical domain.[35] In our experiments, however, the `BidirLSTM` demonstrates clearly stronger performance than the `CNN` counterpart, most likely due to the orders of magnitude larger training data than seen in common text classification tasks, which enables the training of more complex models.

According to the automatic evaluation on the test set (Table 2), `BidirLSTM` is able to suggest a correct subject heading as the top suggestion for 54.35% of the sentences. For the subset of 200 sentences used in the error analysis, the automatic evaluation shows that it is correct for 58.00% of the sentences. However, according to the manual analysis, this number is (at least) 81.50%. One can also assume that the same trend applies when the system is allowed to suggest more than 1 heading per sentence (R@N when $N > 1$). We believe that this points to primarily 2 things that contribute to inconsistency in the data and results in reduced classification performance (according to the automatic evaluation in particular): (1) the dataset spans 2 different documentation standards and 2) the nurses do not always use the correct subject headings when they document (11%-26% incorrect according to the manual error analysis). Interestingly, according to the error analysis, the headings suggested by the `BidirLSTM` method are actually more often correct than the headings originally assigned by the nurses, suggesting that such a model could already be considered for deployment in EHR systems.

Although the `BidirLSTM` model shows surprisingly strong performance in the manual evaluation considering the noisy nature of the training data, state-of-the-art results have also been reported for image classification with noisy training labels,[36] and in fact, noise is intentionally induced in many tasks, such as speech recognition and language modeling, to regularize and improve the generalization of the used models.[37,38] We speculate that a similar effect might be present in our training procedure, leading to a model suggesting more general headings, possibly also preferred by the domain experts in the manual evaluation. Such behavior could be very desirable for standardizing the use of the headings across hospitals and wards, although it leads to deceptively low accuracy scores in the automated evaluation against the noisy data.

As a consequence of noisy data, we have excluded from our original dataset subject headings occurring <100 times. Although the

Table 3. Results from the manual analysis of the originally assigned subject headings and the headings predicted by the best-performing classifier (`BidirLSTM`) for 200 randomly selected sentences.

Class	Original headings	Predicted headings by `BidirLSTM`
1. Correct	0.7400 (148)	0.8150 (163)
2. Maybe correct	0.1500 (30)	0.1300 (26)
3. Incorrect	0.0850 (17)	0.0450 (9)
4. Unable to assess	0.0250 (5)	0.0100 (2)
Automatic evaluation R@1, accuracy (predicted equals original heading)	0.5800 (116)	

Values are decimal (n). The bottom row shows how the classifier performs on these 200 sentences using the R@1 automatic evaluation metric (accuracy).

left out dataset represents only 1%, this cutoff has left out some of the very rare and specialized headings, which can be seen as a study limit. That said, all the methods included in this study were trained and tested on the same dataset.

The experiment indicates that multiple methods are promising for use in sentence classification for nursing notes when trained on a dataset like the one used here. The intended application is a system that can assist nurses in incorporating subject headings when they document. With sentence-level classification, we will be able to assign subject headings to each sentence individually and then structure them into paragraphs based on similarity between their assigned subject headings. In a related study, we have tentatively tested a prototype system with this functionality.[39] There the evaluators reported they believe such a system can be of great help in making nursing documentation in hospitals easier and less time consuming. Moreover, even though the classification models have been trained on semistructured text with manually assigned paragraph-level headings, the prior study shows that such models can be applied on free narratives without any further domain adaptation, although with slightly reduced classification accuracy. We see that a system with this functionality could be particularly useful when the documentation is done through speech-to-text dictation. This will allow nurses to mainly perform the documentation using a microphone while the system will still be able to generate a structured representation with assigned subject headings in an automatic or semiautomatic manner.

Being able to automatically classify arbitrary sentences in clinical nursing notes in terms of their topic or "aboutness" is also useful in other natural language processing tasks like search, information extraction, clustering, automatic template filling, and automatic summarization. Sentence classification may also be beneficial for other health professionals' documentation—who produce similar type of text where some sort of subject headings are being used. Although our experiment is limited to one language (Finnish), the same approach is applicable to other languages as well because the methods are fully data-driven and require no additional language-specific lexical resources.

All neural network approaches tested in our experiments show clear improvement over more traditional methods, such as SVMs, thus warranting further investigation of complex neural models. As future work we will be looking into the latest progress in neural text representations and classifiers, such as ELMo,[40] BERT,[41] and topic memory networks.[35] However, no readily available models for these methods are currently available for clinical Finnish. Further, exploiting the context surrounding each sentence could result in classification improvements, as demonstrated by Song et al.[42]

CONCLUSION

This study explores how 7 text classification methods perform in the task of predicting or suggesting subject headings for individual sentences in nursing notes. According to the automatic evaluation, the results indicate that the best-performing method is the one based on a bidirectional LSTM recurrent neural network architecture. It is able to correctly classify 54.35% of the sentences in the test dataset when it is allowed to suggest 1 subject heading per sentence (R@1, accuracy). When the methods are allowed to suggest 10 subject headings (R@10), the scores are considerably higher where the best method suggests the correct heading in 89.54% of the cases. In addition, a group of domain experts conducted a manual analysis of a data sample which allowed us to compare the headings originally assigned by the nurses and the automated predictions by the best-performing method. This analysis indicates that its actual performance or accuracy is substantially higher than what the automated evaluation score tells, and comparable to the accuracy of the nurses.

We find the results to be encouraging and believe that several of the tested methods can be used to efficiently assist nurses in selecting the most appropriate subject headings when they document administered patient care. A practical impact of such a system would be that nurses save time and effort in tasks related to care documentation, which will result in more time to concentrate on the patient and deliver better care. Another outcome could be improved consistency and correctness in nurses' use of subject headings. Future research should focus on the implementation and extrinsic evaluation of such a system in a clinical setting, including the impact it has on nursing work, quality of documentation, and delivered care.

FUNDING

This work was supported by Business Finland (prev. Tekes, 644/31/2015) and the Academy of Finland (315376).

AUTHOR CONTRIBUTIONS

HM initiated the study and was, together with KH, responsible for the overall design. HM and KH performed all machine learning experiments, data preprocessing, and automated evaluations under the supervision of FG and TS. L-MP, HS, and SS provided the required domain expertise and conducted the manual quality evaluations of the training data and model predictions. HM, KH, and L-MP outlined the first draft of the manuscript. All authors contributed to editing and revising the manuscript. The manuscript has been read and approved by all named authors and there are no other persons who satisfy the criteria for authorship but are not listed. The order of authors listed in the manuscript has been approved by all authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

The authors would like to thank Hanna-Maria Matinolli and Riitta Danielsson-Ojala, who helped them with the data analysis.

CONFLICT OF INTEREST STATEMENT

None declared

REFERENCES

1. Yee T, Needleman J, Pearson M. The influence of integrated electronic medical records and computerized nursing notes on nurses' time spent in documentation. *Comput Inform Nurs* 2012; 30 (6): 287–92.
2. Manor-Shulman O, Beyene J, Frndova H, *et al*. Quantifying the volume of documented clinical information in critical illness. *J Crit Care* 2008; 23 (2): 245–50.
3. Saranto K, Kinnunen U-M, Kivekäs E, *et al*. Impacts of structuring nursing records: a systematic review. *Scand J Caring Sci* 2014; 28 (4): 629–47.
4. Hyppönen H, Saranto K, Vuokko R, *et al*. Impacts of structuring the electronic health record: A systematic review protocol and results of previous reviews. *Int J Med Inform* 2014; 83 (3): 159–69.
5. Häyrinen K, Lammintakanen J, Saranto K. Evaluation of electronic nursing documentation—nursing process model and standardized terminologies as keys to visible and transparent nursing. *Int J Med Inform* 2010; 79 (8): 554–64.
6. Tange HJ, Schouten HC, Kester AD, Hasman A. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *J Am Med Inform Assoc* 1998; 5 (6): 571–82.
7. Forsvik H, Voipio V, Lamminen J, Doupi P, Hyppönen H, Vuokko R. Literature review of patient record structures from the physician's perspective. *J Med Syst* 2017; 41 (2): 29.
8. Rathert C, Mittler JN, Banerjee S, McDaniel J. Patient-centered communication in the era of electronic health records: What does the evidence say? *Patient Educ Couns* 2017; 100 (1): 50–64.
9. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* 2009; 16 (6): 806–15.
10. Li Y, Gorman SL, Elhadad N. Section classification in clinical notes using supervised hidden Markov model. In: *Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10*. 2010: 744–50.
11. Haug PJ, Wu X, Ferraro JP. Developing a section labeler for clinical documents. *AMIA Annu Symp Proc* 2014; 2014: 636–80.
12. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med* 2015; 65 (2): 155–66.
13. Koopman B, Zuccon G, Nguyen A, *et al*. Automatic ICD-10 classification of cancers from free-text death certificates. *Int J Med Inform* 2015; 84 (11): 956–65.
14. Gobbel GT, Reeves R, Jayaramaraja S, *et al*. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. *J Biomed Inform* 2014; 48: 54–65.
15. Uzuner Ö, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008; 15 (1): 14–24.
16. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009; 16 (4): 561–70.
17. Stubbs A, Kotfila C, Xu H, *et al*. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. *J Biomed Inform* 2015; 58: S67–77.
18. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates; 2015: 649–57.
19. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015: 1422–32.
20. Joulin A, Grave E, Bojanowski P, *et al*. Bag of tricks for efficient text classification. *arXiv* 2016 Aug 9 [E-pub ahead of print].
21. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016: 1014–23.
22. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
23. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput* 2000; 12 (10): 2451–71.
24. LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86 (11): 2278–324.
25. Bojanowski P, Grave E, Joulin A, *et al*. Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 2017; 5: 135–46.
26. Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges CJC, Smola AJ, eds. *Advances in Kernel Methods–Support Vector Learning*. Cambridge, MA: MIT Press; 1999, 169–84.
27. Joachims T. Text categorization with support vector machines: learning with many relevant features. In: *European Conference on Machine Learning*. New York: Springer; 1998: 137–42.
28. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 1988; 24 (5): 513–23.
29. Williams CK, Seeger M. Using the Nyström method to speed up kernel machines. In: *NIPS'00 Proceedings of the 13th International Conference on Neural Information Processing Systems*. Cambrige, MA: MIT Press; 2000: 661–7.
30. Rahimi A, Recht B. Random features for large-scale kernel machines. In: *Advances in Neural Information Processing Systems 20*. Red Hook, NY: Curran Associates; 2008: 1177–84.
31. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; 2 (3): 18–22.
32. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2*. Red Hook, NY: Curran Associates; 2013; 3111–9.
33. Harris ZS. Distributional structure. *Word* 1954; 10 (2–3): 146–62.
34. Kaewphan S, Hakala K, Ginter F. UTU: Disease mention recognition and normalization with CRFs and vector space representations. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. 2014: 807–11.
35. Zeng J, Li J, Song Y, *et al*. Topic memory networks for short text classification. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, 3120–3131.
36. Krause J, Sapp B, Howard A, *et al*. The unreasonable effectiveness of noisy data for fine-grained recognition. In: *European Conference on Computer Vision*. Cham, Switzerland: Springer; 2016: 301–20.
37. Hannun A, Case C, Casper J, *et al*. Deep speech: scaling up end-to-end speech recognition. *arXiv* 2014 Dec 19 [E-pub ahead of print].
38. Xie Z, Wang S, Li J, *et al*. Data noising as smoothing in neural network language models. *arXiv* 2017 Mar 7 [E-pub ahead of print].
39. Moen H, Hakala K, Peltonen LM, *et al*. Evaluation of a prototype system that automatically assigns subject headings to nursing narratives using recurrent neural network. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. 2018, 94–100.
40. Peters M, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. 2018: 2227–37.
41. Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* 2019 May 24 [E-pub ahead of print].
42. Song X, Petrak J, Roberts A. A deep neural network sentence level classification method with context information. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018: 900–4.