MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220154

Text Classification Model Explainability for Keyword Extraction - Towards Keyword-Based Summarization of Nursing Care Episodes

Akseli Reunamo [a], Laura-Maria Peltonen [b], Reetta Mustonen [b], Minttu Saari [b,c], Tapio Salakoski [d], Sanna Salanterä [b,e], Hans Moen [f]

[a] Department of Biology, University of Turku, Turku, Finland
[b] Department of Nursing Science, University of Turku, Turku, Finland
[c] Information Management Unit, Satakunta Central Hospital, Pori, Finland
[d] Department of Mathematics and Statistics, University of Turku, Turku, Finland
[e] Turku University Hospital, Turku, Finland
[f] Department of Computing, University of Turku, Turku, Finland

Abstract

Tools to automate the summarization of nursing entries in electronic health records (EHR) have the potential to support healthcare professionals to obtain a rapid overview of a patient's situation when time is limited. This study explores a keyword-based text summarization method for the nursing text that is based on machine learning model explainability for text classification models. This study aims to extract keywords and phrases that provide an intuitive overview of the content in multiple nursing entries in EHRs written during individual patients' care episodes. The proposed keyword extraction method is used to generate keyword summaries from 40 patients' care episodes and its performance is compared to a baseline method based on word embeddings combined with the PageRank method. The two methods were assessed with manual evaluation by three domain experts. The results indicate that it is possible to generate representative keyword summaries from nursing entries in EHRs and our method outperformed the baseline method.

Keywords:

Electronic Health Records; Natural Language Processing; Nursing

Introduction

Nurses document on a daily basis the care provided to each patient undergoing treatment in a hospital. This information is stored in patients' electronic health records (EHRs). The main purpose of EHRs is to ensure optimal care and support the communication between shift personnel, between wards, as well as between hospitals and other healthcare service providers. A large portion of this information is documented as free-text narratives. As large quantities of documented information may accumulate in patients' EHRs, particularly for patients suffering from more complex and long-term health problems, nurses may have insufficient time to study the information previously documented. This is especially the case during busy care situations when time and personnel resources are limited. Consequently, this could lead to communication failures and errors in the provided care [1]. Tools to automate the summarization of content of EHRs could save time in hectic situations and make the patient information more accessible [2–4].

A central challenge in automatic text summarization lies in how to compute relative information importance; how to decide what information should be included in a summary. This is especially challenging when no labeled training data or user queries are available to guide this process. The purpose of this research is to introduce and report on an initial evaluation of a keyword-based summarization method for nursing entries that utilizes the same principles as what is used in machine learning model explainability. The model explanation is also considered a form of explainable artificial intelligence (XAI). This is applied to a text classification model trained on a distantly related proxy task, which is further used to generate a summary containing keywords and key phrases to provide the user with an intuitive overview of the content in the multiple nursing entries written during individual patients' care episodes, meaning we perform keyword extraction and relevance ranking on a multi-document level. A care episode here refers to a patient's stay in the hospital. In brief, this method utilizes a text classification model trained to predict the (Finnish) care classification topic headings used by nurses when they document the care. Each paragraph contains one heading from a taxonomy with more than 400 headings [5,6]. Next, using the classification model we apply a XAI method to extract the most predictive words for the topic heading with the highest confidence. Due to the nature of the prediction task, we make an assumption that words which are seen as most predictive, found using the XAI method, are also the most central keywords associated with the "meaning" expressed in the paragraphs (see the Explainer Extractor section for more details). This provides relevance weights to each word in each paragraph locally. Finally, to rank keywords by their relevance on the level of a care episode, each paragraph is assigned an importance weight that is further multiplied with the normalized keyword weights. As a baseline method to compare against, a keyword extraction method based on PageRank with word embeddings is used (see the PR Extractor section). Three domain experts evaluated the keyword summaries of 40 care episodes extracted by these two methods. We are not aware of previous works reporting on this task.

Related work

Keyword extraction is a task that aims to retrieve the most important words related to the subject of a given text [7]. Keyword extraction methods can be divided into supervised (classification problems where keywords are labels) and unsupervised methods (statistical, entropy-based, and graph-based methods) [8]. Given the nature of the task, most existing methods can be considered unsupervised and domain-independent. A general trend in more recent unsupervised keyword extraction approaches typically is combining multiple

techniques to increase the extraction quality. However, a comprehensive review of such methods is lacking. Zhang et al. [9] combined pre-trained word vectors (trained with word2vec [10]) and TextRank algorithm that used cosine similarity between the word vectors to form the graph for the TextRank model. This method outperformed the use of TF-IDF as word representation, as well as the original TextRank and UNT-TextRank, at the task of extracting keywords from computer science-related literature. Teneva et al. [11] combined PageRank and Latent Dirichlet Allocation (LDA) for keyword extraction (Salience Rank) where LDA and topic and corpus specificity were used to form the graph for the PageRank algorithm. This method outperformed topical PageRank and single topical PageRank in keyword extraction from the 500N-KPCrowd and Inspec datasets. Chengzhang and Dan [12] combined PageRank and word2vec word vectors to rank sentences that outperformed TextRank and TF-IDF in Chinese news summarization.

Supervised methods have gained popularity in keyword extraction due to their good performance. However, the complexity of the task and the need, and the lack of tailored training data limits their generalizability [8]. Tang et al. [13] used a BERT-based model with an added attention layer to extract keywords based on attention layer weights from clinical notes. They report that the attention-based model can identify relevant keywords that are strongly related to the clinical progress note categories. However, the quality of the extracted keywords was not assessed. In addition, attention-based interpretability has been found to be inconsistent with model predictions and attention does not necessarily correspond to the importance of input for prediction [14,15]. Meng et al. [16] used a RNN-based supervised generative model to predict keywords from multiple scientific publication datasets which outperformed TF-IDF, TextRank, SingleRank, ExpandRank, Maui, and KEA. In addition, multiple more shallow supervised methods such as Naive Bayes [17] and Random Forest [18] have been used in keyword extraction, but their performance only exceeds very basic benchmark algorithms. Our work is comparable to that of Tang et al. [13] who performed keyword extraction from clinical text using an attention-based model. However, we utilize the local interpretable model-agnostic explanations (LIME) [19] to extract keywords from our classification model and focus on multi-document keyword extraction.

Methods

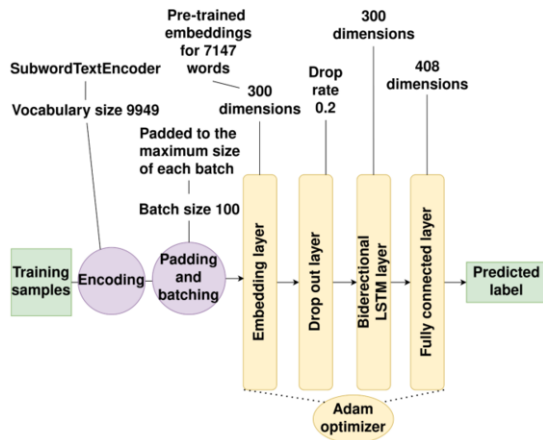
Generation of keyword-based summaries was performed with two methods: Explainer Extractor and PR Extractor. PR Extractor was chosen as the baseline because it enhances the classic PageRank algorithm by incorporating better semantic representation which seems to perform better than the plain method.

Explainer Extractor - Keyword extraction through explainable AI

This method combines a text classification model with model explainability. Model explainability is used to extract the most relevant words for the classification task. For the text classification, we use a bidirectional LSTM-based neural network [20], and for the model explainability, we use local interpretable model-agnostic explanations (LIME) [19].

The classification model was trained to classify each paragraph into one subject heading (multiclass classification). The bidirectional LSTM layer structure and hyperparameters were chosen based on its performance in previous research [6]. The full pipeline of training the neural network is presented in Figure 1. Tensorflow (version 2.2.0) was used to implement the model [21].

Figure 1– The model architecture and training pipeline. Each sample was encoded with TensorFlow datasets module SubwordTextEncoder to numerical format. After that samples were batched and padded and fed to the model. The Embedding layer used pre-trained word embeddings from Word2Vec (see section PR Extractor for details) if one for the word was available. Early stopping with the patience of three epochs was used to avoid overfitting.



Data

The used dataset contains nursing entries obtained from a Finnish university hospital. Each entry was split into paragraphs and corresponding subject heading. The whole dataset consisted of 1,7M paragraphs and 408 unique heading pairs. Each paragraph-heading pair formed individual training examples. The dataset was split into a training set (60%), a validation set (20%), and a test set (20%).

Keyword extraction

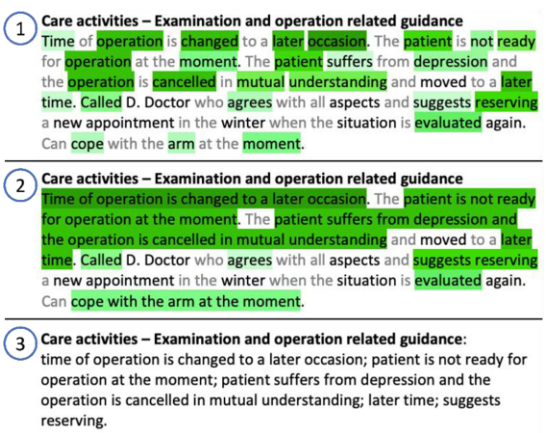
After training the base model, the LIME package's module LimeTextExplainer was used to get the local importance of words in each paragraph. LimeTextExplainer uses Ridge regression to determine the regression coefficients for each word in a sample and the words which get the highest absolute coefficient values are interpreted as the most important ones. LimeTextExplainer leverages the base model to investigate how perturbations in original input affect the prediction [19]. To rank the paragraphs by their importance, a paragraph-similarity graph was constructed; paragraphs were nodes, and edges were weights between their similarities. These edge weights were calculated using the text representation from the bidirectional-LSTM layer as a vector representation of each paragraph, and the cosine similarity between paragraph-vector pairs was used to construct the edge weights. By applying the PageRank algorithm [22], paragraphs were ranked according to their importance.

The full keyword extraction pipeline has the following steps.

1. The care episode was split into paragraph and heading pairs, and only alphabetic characters were retained in each paragraph.
2. Each paragraph of each care episode was ranked according to its importance.
3. Keywords were extracted from each paragraph with LimeTextExplainer and coefficients with the highest absolute values were used as keywords.

4. Coefficients were Z-standardized and further weighted with paragraph score. If a paragraph had only one token, the importance of a word was assigned to be the same as the importance of the paragraph.
5. Keywords occurring next to each other were combined into keyphrases and the highest scores of the components were used as keyphrase scores (Figure 2).
6. Stopwords defined in nltk.stopwords('finnish') [23] and some additional common Finnish stop words were removed.
7. Keywords were mapped back to corresponding headings.
8. Duplicate keywords were removed.
9. The number of keywords was 10% of care episode tokens which were returned in order of their original appearance in the care episode.

Figure 2- Example of word scoring. This example paragraph is translated from Finnish to English. 1) The importance of keywords is in relation to the darkness of the token's background color from grey to dark green. 2) Keywords occurring next to each are combined into keyphrases. 3) Illustration of how a keyword summary with the top five keywords and keyphrases generated from this paragraph alone would look like.



PR Extractor - Keyword extraction with modified PageRank

This method uses a modified version of the PageRank algorithm. It uses word embeddings, and their cosine similarity scores to form a word-similarity graph before applying the PageRank algorithm.

Base model and data

The word2vec toolkit was used to obtain vector representation of words [10]. The model was trained with the skip-gram architecture, a dimensionality of 300, and otherwise default hyperparameters. Corpus used for training contained 136M tokens obtained from nursing and doctors' entries from the same hospital as above.

Keyword extraction

After training, word similarities were calculated using cosine similarity. These were used to create a word-similarity graph before applying the PageRank algorithm. The algorithm calculates word importance based on their centrality in the graph.

The following is the full keyword extraction pipeline for a care episode:

1. Keywords were extracted from the full care episode.
2. Stopwords defined in nltk.stopwords('finnish') and some additional common Finnish stop words were removed.
3. Neighboring keywords were combined into keyphrases and the highest scores of the components were used as phrase scores.
4. Duplicate keywords were removed.
5. Keywords were mapped back to corresponding headings.
6. 10% of care episode tokens were returned in order of their appearance in the care episode.

Experimental setup

Keyword summaries were extracted using the two methods from 40 randomly selected care episodes that were not used in training. The care episodes consisted of at least 5 and at most 15 individual nursing entries. The evaluation was then performed by three domain experts with a nursing background who we here refer to as evaluators. First, two of the evaluators evaluated the keyword summaries of both methods independently. Finally, the third one assessed and decided on a consensus for the disagreeing assessments. For each care episode, the evaluators were instructed to read the keyword summary, then read all the original nursing entries from the care episodes, and finally return to the keyword summary to score it. The main question for the evaluators was how well the keyword summaries succeed at conveying the necessary information about the documented nursing care in a way that it enables them to provide an intuitive overview of the content of the care episodes. Each keyword summary was evaluated with a four-class rating (Table 1).

Statistical significance between the ratings of the two methods was tested with Wilcoxon's signed-rank test. A two-tailed 0.5 point effect with an alpha of 0.05 and a power of 0.80 yielded a needed sample size of 35 to detect the effect. A total of 40 care episodes were evaluated to ensure a sufficient sample size. The manual evaluation and statistical analysis were blinded to which method was the Explainer Extractor and which was the PR Extractor.

Table 1- Manual evaluation scale.

Rating	Explanation
4	This adequately conveys an overview of the information in the care episode.
3	This only partly conveys an overview of the information - central information/keywords are missing
2	This poorly conveys the information.
1	Unable to assess.

Results

The Bidirectional-LSTM model used as the backend of Explainer Extractor had a prediction accuracy of 73.6% on the test set.

The results of the manual evaluation are presented in Table 2. Explainer Extractor achieved the best results with 55% of the summaries being rated as adequate for conveying an overview of the information in the associated care episodes (rating of 4). For PR Extractor, 22.5% of the summaries got a rating of 4. All

keyword summaries extracted by both methods could be assessed. The median for the evaluations of the Explainer Extractor was 4 (IQR 1) and the median for the PR Extractor was 3 (IQR 0). A Wilcoxon signed-rank test showed a statistically significant difference between the ratings of the two methods ($Z = -3.021$, $p = 0.003$).

Table 2 - Results of the manual evaluation.

Rating	Explainer Extractor	PR Extractor
4	55.0 % (22)	22.5 % (9)
3	42.5 % (17)	67.5 % (27)
2	2.5 % (1)	10.0 % (4)
1	0.0 % (0)	0.0 % (0)

Discussion

The results indicate that it is possible to make keyword-based summaries from care episodes, where the best performing method, Explainer Extractor, generated summaries rated as adequate for 55% of the care episodes (rating of 4). Only 2.5% of the summaries poorly conveyed the information in the associated care episodes. In the future, this type of automated summary could be useful during emergency situations when a patient is rapidly deteriorating and there's a need to get a quick glance of the patient's health history.

Previous work has shown that supervised keyword extraction methods, based on tailored training data, can provide good performance [8]. However, the task specificity reflected in the training data also puts restrictions on their applications. The approach presented here relies instead on a very different type of training data that is readily available from the original data and can thus be seen as a task-agnostic approach to keyword extraction since it does not need tailored training data for this exact purpose.

Future research is needed to gain knowledge on how these methods, Explainer Extractor in particular, can be further improved by exploring better ways to rank paragraphs and keywords and how to represent them better for the user. We are planning to explore how the use of LIME for keyword extraction and summarization performs for other data sets, where other "proxy" classification tasks can be formulated, for example, physician notes with ICD codes, or clinical progress note categories (as in [13]). In addition, other text classification models, such as the more recent transformer-based models, as well as other explanation methods for extracting keywords and key phrases may increase the quality of keyword summaries. In the reported experiment, summaries were presented to the evaluators simply as a set of boxes, one for each unique topic heading and the associated keywords and key phrases (see Figure 2 part 3).

As future research, we also plan to explore other ways of visualizing such summaries to the user, e.g., through a timeline-based interface. An interactive visualization system could for example enable users to see full sentences, paragraphs, or documents when more detailed information is needed. Displaying the full text to the user with keywords and key phrases simply highlighted (as in Figure 2 part 2) might be another way of enhancing reading speed.

Conclusions

We tested two methods at the task of automatically generating keyword summaries from nursing entries in patient care episodes. We found that our proposed method based on model explainability, or XAI, outperformed a baseline method that relied on word embeddings and PageRank. The results are promising and indicate that this method could be helpful to nurses in providing them with an intuitive overview of the information documented in patient care episodes, particularly when time is limited.

Acknowledgements

The study had an ethical review and hospital administrative approval. The authors state no conflict of interest. The research was supported by the Academy of Finland (315376).

References

[1] H. Alfredsdottir, and K. Bjornsdottir, Nursing and patient safety in the operating room., *J. Adv. Nurs.* **61** (2008) 29–37.

[2] T.T. Van Vleck, D.M. Stein, P.D. Stetson, and S.B. Johnson, Assessing data relevance for automated generation of a clinical summary., *AMIA Annu. Symp. Proc.* (2007) 761–765.

[3] H. Moen, L.-M. Peltonen, J. Heimonen, A. Airola, T. Pahikkala, T. Salakoski, and S. Salanterä, Comparison of automatic summarisation methods for clinical free text notes., *Artif Intell Med.* **67** (2016) 25–37.

[4] R. Pivovarov, and N. Elhadad, Automated methods for the summarization of electronic health records., *J. Am. Med. Inform. Assoc.* **22** (2015) 938–947.

[5] P. Hoffrén, K. Leivonen, and M. Miettinen, Nursing Standardized Documentation in Kuopio University Hospital, in: IOS Press, 2009: pp. 776 – 777.

[6] H. Moen, K. Hakala, L.-M. Peltonen, H. Suhonen, F. Ginter, T. Salakoski, and S. Salanterä, Supporting the use of standardized nursing terminologies with automatic subject heading prediction: a comparison of sentence-level text classification methods., *J. Am. Med. Inform. Assoc.* **27** (2020) 81–88.

[7] G.K. Palshikar, Keyword Extraction from a Single Document Using Centrality Measures, in: A. Ghosh, R.K. De, and S.K. Pal (Eds.), Pattern recognition and machine intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 503–510.

[8] N. Firoozeh, A. Nazarenko, F. Alizon, and B. Daille, Keyword extraction: Issues and methods, *Nat. Lang. Eng.* **26** (2020) 259–291.

[9] Y. Zhang, F. Chen, W. Zhang, H. Zuo, and F. Yu, Keywords extraction based on word2vec and textrank, in: Proceedings of the 2020 The 3rd International Conference on Big Data and Education, ACM, New York, NY, USA, 2020: pp. 37–42.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, *arXiv*. (2013).

[11] N. Teneva, and W. Cheng, Salience Rank: Efficient Keyphrase Extraction with Topic Modeling, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2017: pp. 530–535.

[12] X. Chengzhang, and L. Dan, Chinese text summarization algorithm based on word2vec, *J. Phys.: Conf. Ser.* **976** (2018) 012006.

[13] M. Tang, P. Gandhi, M.A. Kabir, C. Zou, J. Blakey, and X. Luo, Progress Notes Classification and Keyword Extraction using Attention-based Deep Learning Models with BERT, *arXiv*. (2019).

[14] S. Serrano, and N.A. Smith, Is Attention Interpretable?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019: pp. 2931–2951.

[15] S. Jain, and B.C. Wallace, Attention is not Explanation, *arXiv*. (2019).

[16] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, and Y. Chi, Deep Keyphrase Generation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2017: pp. 582–592.

[17] C. Caragea, F.A. Bulgarov, A. Godea, and S. Das Gollapalli, Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Stroudsburg, PA, USA, 2014: pp. 1435–1446.

[18] A.K. John, L. Di Caro, and G. Boella, A supervised keyphrase extraction system, in: A. Fensel, A. Zaveri, S. Hellmann, and T. Pellegrini (Eds.), Proceedings of the 12th International Conference on Semantic Systems - SEMANTiCS 2016, ACM Press, New York, New York, USA, 2016: pp. 57–62.

[19] M.T. Ribeiro, S. Singh, and C. Guestrin, Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, New York, USA, 2016: pp. 1135–1144.

[20] S. Hochreiter, and J. Schmidhuber, Long short-term memory., *Neural Comput.* **9** (1997) 1735–1780.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, and G. Brain, TensorFlow: A System for Large-Scale' ' Machine Learning, (2016).

[22] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Stanford InfoLab, 1999.

[23] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit, 1st ed., O'Reilly Media, Beijing, 2009.

Address for correspondence

Akseli Reunamo, akseli.y.reunamo(at)utu.fi.