

A Document Tagging Support System for Nursing Care Experts^{*}

Beat Tödli^{1,*}, Sebastian Müller¹, Melanie Rickenmann¹, Janine Vetsch¹ and Simon Haug¹

¹Eastern Switzerland University of Applied Sciences, Rosenbergstrasse 59, 9000 St. Gallen, Switzerland

Abstract

We present the findings of an interdisciplinary project that implemented a document tagging support system for nursing care experts. We evaluate its performance and provide lessons learned. This project was particularly marked by a low inter-rater reliability of the document labels and use case understanding issues, but also of a good performance of a simple, BERT-based binary relevance approach.

Keywords

document tagging support, document classification, inter-rater reliability, nursing care professional education

1. Introduction

Text mining and document classification are long-standing research areas [1] where deep neural networks have made very significant contributions over the past years. In particular, bidirectional transformer models such as BERT seem to "learn" structural information about language [2] and provide unprecedented performance and ease of use [3]. Such models can be used for a broad range of application classes, such as document classification [4], regression tasks, document tagging, information retrieval, recommendation tasks, and many more [5]. While this is ground-breaking, it opens up a wide range of applied machine learning research opportunities. The challenge there lies in gathering, structuring, consolidating and spreading experiences into domain-adapted methodologies and insights. Peculiar challenges are often raised by concrete application cases, such as in the case study reported here.

We present the findings of a small-scale applied interdisciplinary project in the domain of tagging support for document labelling tasks. The project's goal was building a document tagging support system for health experts tasked with tagging nursing care publications. As an applied machine learning project, it had a set of requirements and challenges that are quite different from standard text classification or information filtering tasks, on which we report here.

LWDA'23: Lernen, Wissen, Daten, Analysen. October 09–11, 2023, Marburg, Germany


*Corresponding author.

✉ beat.toedtli@ost.ch (B. Tödli)

🌐 <https://www.ost.ch/de/person/beat-toedtli-1039> (B. Tödli); <https://www.ost.ch/de/person/sebastian-mueller-940> (S. Müller); <https://www.ost.ch/de/person/melanie-rickenmann-1090> (M. Rickenmann); <https://www.ost.ch/de/person/janine-vetsch-1046> (J. Vetsch)

🆔 0000-0003-3674-2340 (B. Tödli); 0000-0001-9877-0086 (S. Müller); 0009-0004-7421-0905 (S. Haug)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. The Application Case

2.1. Business Understanding

Researchers in nursing care are investigating how to promote continuous professional learning in the daily practice of nursing [6]. In performing patient documentation, a tool could integrate scientific evidence at the point of care and therefore provide nurses with easy access to evidence in daily work. They therefore are labelling a dataset of nursing care publications and are associating one or several topic tags with each of them¹. This report is concerned with building a document tagging support system for these experts, to be used as a decision support aid.

In the business understanding efforts it was realized rather late in the project that helping experts to save on the time needed to inspect a document was not as relevant as motivating the expert to rethink his or her tagging decisions.

2.2. Data Understanding

The data consisted of 1515 nursing care or medical publications each associated with one or several of 24 different tags. The number of publications per tag (see Fig. 1) arose historically and reflects a data taking campaign that did not specify the relative frequency of the tag categories in the data set. After dropping tags with less than 20 associated documents and the class others, 1293 documents in 18 relevant categories remained.

The tag frequency distribution was a data understanding indicator that had major implications on the project: Only few tags were associated per document, but a significant fraction of documents had more than one tag associated with it. 76% of all documents had 1 tag, 23% had two tags and 1% had three or more tags.

Therefore, a multi-label classification approach [7] is adequate since in 24% of the cases more than one tag is associated with a document.

Furthermore, metrics such as $\text{precision}@k$ need to be evaluated with respect to the multi-label case. In a UX workshop with the experts it was found desirable to present $k = 3$ selected document tags. This implies that $\text{precision}@3$ values cannot reach a value of 100%. Also, most information retrieval or recommender systems have a much smaller percentage of relevant items to retrieve or recommend, so that our precision values will likely be much higher.

2.3. Dealing with Imprecise Labels

A complication arose with the early realisation that there was some disagreement between experts on which labels to assign to a given document. This effect was expected based on the results of Xia and Yetisgen-Yildiz, since medical training alone does not ensure high inter-annotator agreement and no NLP researcher had been involved in the annotation process until this project. [8] The inter-expert labelling reliability was assessed using a small labelling campaign where two experts labelled the same 60 documents. The co-occurrence matrix of the ratings as judged by two experts indicated disagreements even though for most documents (87%) at least one tag was overlapping. It was also found that averaged over all assigned tags, the

¹We use the word "tag" or "category" instead of "label" to indicate that each document can have more than one tag (or category) associated with it.

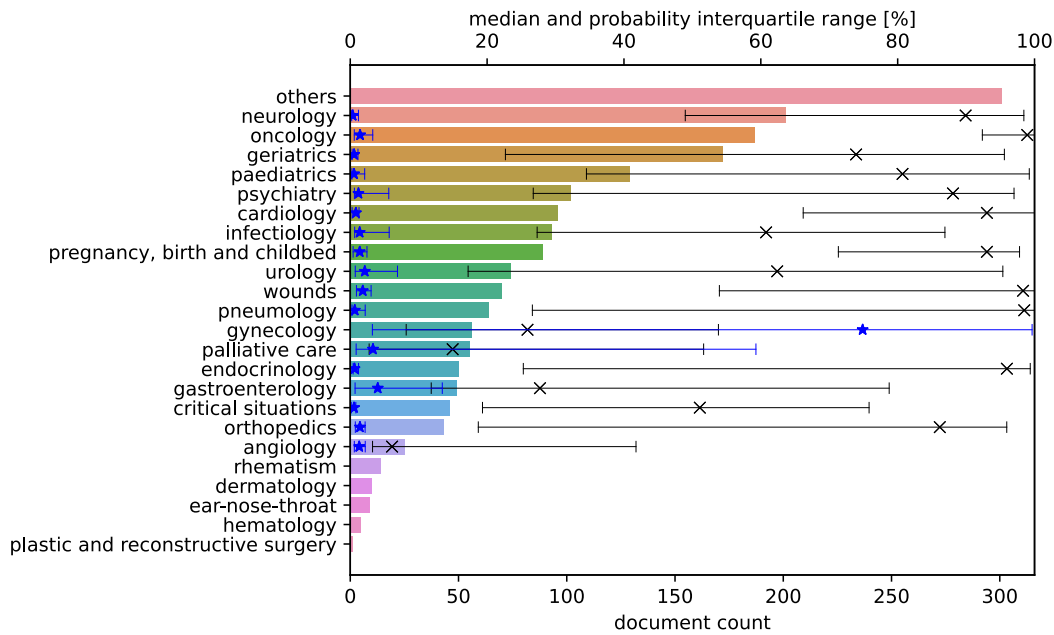


Figure 1: The number of documents per tag category (bars, lower x-axis) and the median and interquartile range of the predicted (test) class probabilities ('x'- and '*'-marked points with bars, upper x-axis). The black, 'x'-marked bars correspond to documents containing the given tag, the blue, star-marked bars to documents not containing the given tag. Documents associated with only the label "others" and labels with less than 20 associated documents were not used for training. Noticeably, more training data is generally helpful.

intersection over union of the assigned tag sets was 72%. However, to assess inter-expert reliability, Krippendorff's alpha [9] is more appropriate here, as it takes into account interlabeller (dis-)agreement by chance. It has further advantages in that it allows for missing values and multiple tags. We find a value of $\alpha = 0.59$, indicating a substantial inter-rater agreement according to Landis and Koch [10], but not according to Krippendorff who is reported to consider 0.8 as an absolute minimum value for any serious purpose [11]. Based on these indications, an important task in a next iteration is to write annotation guidelines and ensure consensus between the various nursing care experts about how to apply these guidelines [8].

3. Functional Prototype Construction

3.1. Data Preprocessing and Feature Engineering

The proposed system consists of text extraction and feature engineering steps that return the BERT sentence embedding using the Hugging Face model "paraphrase-MiniLM-L6-v2"² of the paper abstract and a tag filtering step based on this vector to be detailed further in Sec. 3.2.

²See <https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

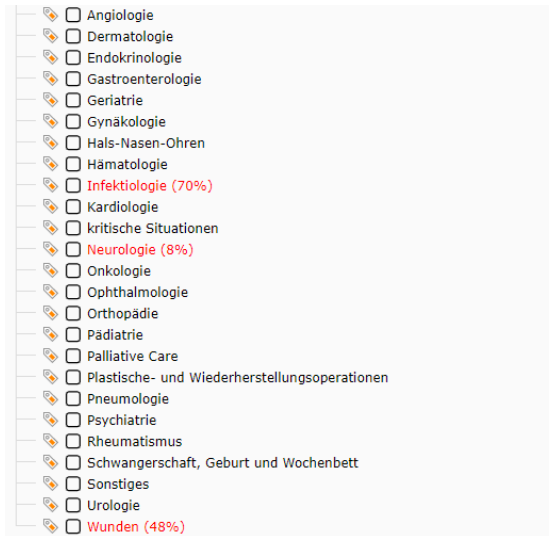


Figure 2: Screenshot of the application’s tagging interface, where the three filtered tags are highlighted in red. The predicted tag probabilities are given in brackets.

A feature selection decision was made by using BERT feature vectors of the publication abstracts only. They were extracted manually by the experts and have the advantage of fitting well within the 512 subword tokens limit imposed by BERT.

3.2. Modelling: Multi-Label Classification for Document Tag Filtering

For multi-label classification, we used a binary relevance problem transformation method [12]. Let $n_c = 18$ be the number of different tags available before filtering. Binary probabilistic classifiers $f_i : X \rightarrow [0, 1]$ for each $i \in \{1 \dots n_c\}$ were trained using support vector classifiers with rbf-kernels. Their predicted class probabilities for a document x , $\{f_i(x), i = 1, \dots n_c\}$, were interpreted as tag relevance scores. The tags with the $k = 3$ highest scores were selected.

As part of the user interface design choice and as a consequence of the distribution of the number of tags per document, $k = 3$ tags were selected and presented in red with the associated probabilities. Fig. 2 shows the user interface as it is currently implemented.

3.3. Evaluation: Statistical Results

We discuss the evaluation of the data mining prototype on the statistical level. The user level evaluation is discussed in 4.2. Our main results are given in Tab. 1

The low precision@3 and MAP@3 values can be attributed to the fact that the number of relevant tags per document are mostly 1 or 2. The average precision@1 is higher, at 82%. From a practical point of view, the recall@3 value is presumably the most important one, as three tags can easily and quickly be judged by an expert. By looking only at these top-3 selections, however, the expert might miss relevant tags in 8% of the cases.

This result was deemed sufficient for a first iteration and therefore terminated classifier optimization efforts, as early user experience feedback was deemed more important. However,

Precision@3	42%
Precision@1	82%
Recall@3	92%
MAP@3	61%

Table 1

Precision, recall and mean average precision@3 values of the SVM-rbf tagging support system averaged over all documents in the test set. These correspond to the best obtained filtering model selected on the validation set.

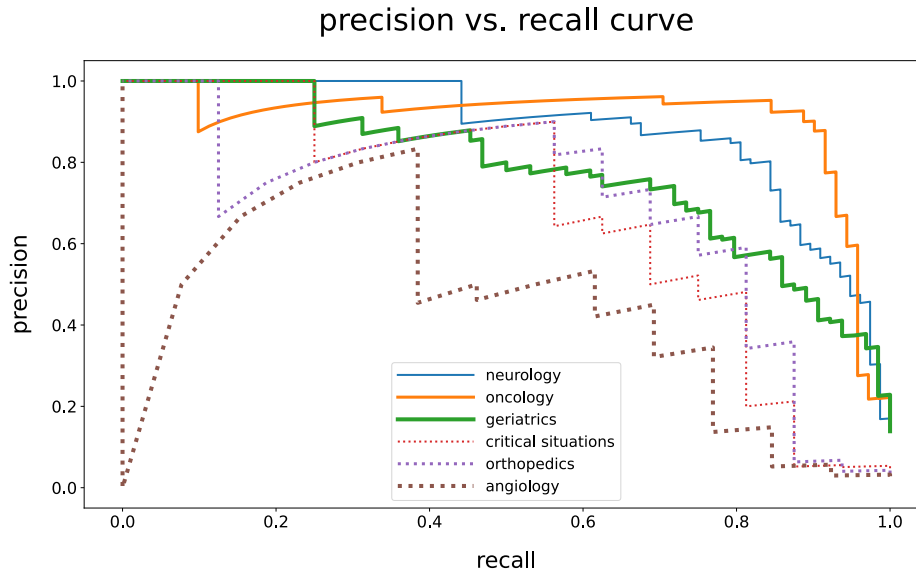


Figure 3: Precision-Recall curves of the single tag relevance score estimators for the 3 most frequent tags (neurology, oncology, geriatrics - solid lines) and the 3 least frequent tags (critical situations, orthopedics, angiology - dotted lines).

the single-tag binary classifiers were far from perfect, as the precision-recall curves in Fig. 3 show. In fact, these curves show that the less frequent tags (with a prevalence of 2-4%) perform worse than the more frequent ones (appearing in 6-13% of the documents), suggesting that a larger data set will likely help improve the performance further.

Although useful as a rough performance estimate, recall@3 must be considered a flawed metric since the number of relevant tags varies mostly between 1 and 2 tags per document. Given that $\sim 97\%$ of all documents in the test data set had at most two associated labels, Tab. 2 considers the documents with one and two labels separately.

On the subset of $\sim 63\%$ of documents with only one relevant tag, Tab. 2 lists the percentages of documents for which this relevant tag is not found, found at the first, second or third item in the filtered list. For the $\sim 33\%$ of documents with two relevant labels, similarly the percentages of where the two relevant tags were found are given.

Thus, for documents with one relevant tag, precision@3 values of over 90% were reached,

Classifier Type \ Position	1 relevant tag				2 relevant tags				
	None	1	2	3	None	1 tag	1+2	1+3	2+3
Support Vector Machine	7%	80%	10%	3%	1%	18%	64%	12%	5%
Naïve Bayes	8%	76%	13%	3%	3%	20%	61%	7%	7%

Table 2

Position of the relevant tags in the filtered list for documents with one and two relevant tags, respectively. Here "1+2" is to be read as "at the first and second position", "1 tag" as "one relevant tag found", and "None" as "no relevant tag selected".

whereas for documents with two relevant tags, both tags were found in the top 3 recommendations in at least 92% of the cases. Using a naïve Bayes classifier instead of a support vector machine generally resulted in a minor performance reduction.

4. Deployment

4.1. Software Architecture

All tagged publications are made accessible for nursing practitioners on a TYPO3-based website. Data is stored on a Linux server and managed in a MariaDB database.

Experts administrate publications in the backend of the website. The Python-based filter system is started once a day with a cron job, retrieves newly uploaded publications via REST-API and writes the filtered tags (also via REST-API) into a designated table of the MariaDB database, from where they are displayed to the tagging experts (see Fig. 2).

4.2. Evaluation: Assessment of User Experience

The system has so far been in use for 6 months. Based on an interview conducted with the nursing care expert chiefly tasked with tagging documents, several crucial insights could be established. The most problematic one was that the expert viewed the filtered tags as almost authoritative recommendations. The expert looked at the system's three filtered tags and checked their plausibility, instead of looking at the document and determining the relevant tags. The option of tagging the document with a non-filtered item and the tag probability indications were ignored. This was done mainly out of convenience and efficiency, believing that no tag could be relevant that was not highlighted. The expert also declared that he no longer looked at the abstract and/or the paper, instead he would rely solely on the filtering system and the title.

These practices must be registered with alarm, since apparently even with usage instructions and being involved in this project and knowledge of the statistical evaluation results, the expert seemed happy to uncritically pass on the responsibility for the correctness of the tags to the tag filtering system.

5. Conclusion and Outlook

We built a document tagging support system to aid nursing care experts in building a labelled dataset for on-the-job professional education.

Using a simple BERT-based feature engineering approach combined with a standard radial basis function support vector machine, recall@3 values of around 90% were reached. While the domain experts deemed this result good enough for deployment, issues regarding the specified inter-rater reliability and scarce training data for several tags show potential for improvements. Consequently, Fig. 1 showed that not all tags were reliably recognized by the system. It is likely that inter-labeller agreement was the limiting factor in this regard, in contrast to many CRISP-DM-projects where data collection, feature engineering and optimizing the modelling step often pose the key challenges.

Our results also demonstrate that building custom tagging support systems is already quite inexpensive. This again suggests that domain-adaptation efforts are reasonably likely to be successful when transformer-based sentence embeddings are used.

We cautiously try generalise these insights. When trying to judge the probability of success of a tagging support system, some important positive indications are if

- the domain-specific dataset is at least moderately sized. In the case of publications, 20-50 abstracts per category might suffice.
- the labels are easily distinguished based on the text *only*, possibly even by non-experts.
- experts judge the reasoning necessary to distinguish between categories to be simple.

During our project, we designed the system with the the goal of challenging the experts to reconsider their tag selections with the aid of the tagging support system. The system was therefore designed around assisting expert labelling, rather than providing an expert opinion by itself. Despite respective efforts and instructions though, the expert interviewed after using the system for two months developed a significant inclination to uncritically adopt the system's (alleged) "recommendation" out of convenience. This UX-challenge remains unsolved and offers great potential for future research.

5.1. Lessons to Consider

In summary, we list some lessons learned:

- Carefully think about how the system biases the expert's labelling decision. The experts might need training to correctly use the system. Otherwise there is a possibility that they uncritically accept the system's biases in the tagging process by trying to become more efficient. Additionally regularly check the way it is used after deployment.
- Do not start a tagging or labelling process without defining clear labelling instructions.
- Monitor the label correlations in the labelling process, and discuss label categories among experts to verify that they are discernible by all [12, 7].
- If in doubt, measure the inter-rater reliability early on. A low inter-rater reliability may invalidate the project goals, and the system performance will be limited by the label consistency the training data has.
- For the task of scientific document tagging support, abstracts are useful. The bulk of a document should be considered in a second iteration only because the limited transformer input length poses additional potentially expensive challenges.

References

- [1] A. Bilski, A review of artificial intelligence algorithms in document classification, *International Journal of Electronics and Telecommunications* vol. 57 (2011). URL: <http://journals.pan.pl/Content/86895/PDF/35.pdf>. doi:10.2478/v10177-011-0035-6.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc., 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- [3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://aclanthology.org/2020.emnlp-demos.6>. doi:10.18653/v1/2020.emnlp-demos.6.
- [4] A. Adhikari, A. Ram, R. Tang, J. Lin, Docbert: Bert for document classification, *ArXiv abs/1904.08398* (2019). URL: <http://arxiv.org/abs/1904.08398>.
- [5] R. Nogueira, K. Cho, Passage re-ranking with bert, 2019. URL: <https://arxiv.org/abs/1901.04085>. doi:10.48550/ARXIV.1901.04085.
- [6] R. Ranegger, S. Haug, J. Vetsch, D. Baumberger, R. Bürgin, Providing evidence-based knowledge on nursing interventions at the point of care: findings from a mapping project, *BMC Medical Informatics and Decision Making* 22 (2022) 308. URL: <https://doi.org/10.1186/s12911-022-02053-8>. doi:10.1186/s12911-022-02053-8.
- [7] R. B. Pereira, A. Plastino, B. Zadrozny, L. H. Merschmann, Correlation analysis of performance measures for multi-label classification, *Information Processing & Management* 54 (2018) 359–369. URL: <https://www.sciencedirect.com/science/article/pii/S0306457318300165>. doi:<https://doi.org/10.1016/j.ipm.2018.01.002>.
- [8] F. Xia, M. Yetisgen-Yildiz, Clinical corpus annotation: challenges and strategies, in: *Proceedings of the third workshop on building and evaluating resources for biomedical text mining (BioTxtM'2012) in conjunction with the international conference on language resources and evaluation (LREC)*, Istanbul, Turkey, 2012, pp. 21–27.
- [9] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology* (second edition), Sage Publications, 2004.
- [10] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, *Biometrics* 33 (1977) 159–174. URL: <http://www.jstor.org/stable/2529310>.
- [11] R. Artstein, M. Poesio, Survey article: Inter-coder agreement for computational linguistics, *Computational Linguistics* 34 (2008) 555–596. URL: <https://aclanthology.org/J08-4004>. doi:10.1162/coli.07-034-R2.
- [12] J. Read, A. Bifet, G. Holmes, B. Pfahringer, Scalable and efficient multi-label classification for evolving data streams, *Machine Learning* 88 (2012) 243–272. URL: <https://doi.org/10.1007/s10994-012-5279-6>. doi:10.1007/s10994-012-5279-6.