Leveraging Machine Learning and Natural Language Processing for Monitoring E-health Publications

Andrius Budrionis, Rune Pedersen, Torbjørn Torsvik, Karianne Lind, Omid Saadatfard
Norwegian Centre for E-health Research
University Hospital of North Norway
Tromsø, Norway
e-mail: Andrius.Budrionis@ehealthresearch.no, Rune.Pedersen@ehealthresearch.no, Torbjorn.Torsvik@ehealthresearch.no,
Karianne.Lind@ehealthresearch.no, Omid.Saadatfard@ehealthresearch.no

***Abstract*—E-health is a rapidly developing field governed by national and international guidelines. These guidelines are often generic, making it problematic to monitor the development of the field with regards to the expected directions. To address this shortcoming, we present a data analytics pipeline for continuous monitoring of e-health publications in Norway with regards to the national e-health strategy. The pipeline contains PubMed data import module, machine learning, natural language processing modules and a visualization component. The potential of the proposed approach is illustrated by identifying publication trends in Norway for the last ten years. These trends show how well focus areas of the Norwegian e-health strategy are represented in scientific publications. The pipeline is customizable and can be extended to support other countries, e-health strategies and publication channels.**

Keywords-e-health strategy; publications; machine learning; natural languge processing.

I. INTRODUCTION

In recent years, e-health has become an important topic in political agendas of both developing and developed countries. According to the World Health Organization, 58% of member states have developed national e-health strategies and 87% reported one or more national mobile health (m-health) initiatives in 2016 [1]. These initiatives are often connected to ongoing research activities disseminating achievements and lessons learned through scientific publication channels.

In Norway, e-health has undergone a strong national foundation process, especially from 2016 up till now. Through the establishment of a Directorate for E-health (E-Dir) and the Norwegian Center for E-health Research (NSE), as well as national initiatives and strategies, such as One Citizen – One Journal [2], the National E-health Strategy [3], and the National Action Plan for E-health 2020-22 [4], the national development and focus on e-health has escalated. In light of these initiatives, monitoring how well national development of e-health corresponds to the goals defined in the aforementioned documents is an important feedback to the decisionmakers.

Considering the multifaceted nature of e-health [5], it is difficult to draw a boundary between e-health, telemedicine, health technology, medicine and other closely related fields. This uncertainty makes it difficult to isolate and measure achievements within e-health and, thus, monitor compliance of national e-health development against strategic documents, such as National E-health Strategy [3] and the National Action Plan for E-health 2020-22 [4].

Publication of scientific papers could be considered as a proxy for research and development in the field [6]. Building on our previous publication [7] where we manually searched for e-health papers and classified them into several groups, we present improvements allowing continuous monitoring of scientific publications in Norway with regards to the National E-health Strategy [3].

Our previous work showed that it is problematic to classify publications into the focus areas of the National E-health Strategy [3] with high accuracy, however, achieved performance was considered sufficient for revealing publication trends. These trends showed value without being completely accurate at a single publication level [7]. This paper presents continuation of our work on e-health publication monitoring in Norway and classification of the identified manuscripts into the six underlying focus areas: 1) digitization of work processes, 2) seamless/coherent patient pathways, 3) improved use of health data, 4) new ways to provide healthcare, 5) common foundation for digital services and 6) national e-health management and increased implementation [3].

The reminder of this paper is organized as follows. Section II provides a summary of methods used to produce results, which are presented in Section III. Section IV discusses the key findings and limitations of this work, while Section V concludes the paper.

II. METHOD

To provide continuous monitoring of e-health publications, a data analytics pipeline, covering data collection analysis and visualization was developed. Considering the multifaceted definition of e-health, a neural network-based approach, encompassing language representation was selected for differentiating e-health publications from other irrelevant content. To be specific, a Bidirectional Encoder Representations from Transformers (BERT-base) model pretrained on large Wikipedia corpus [8] was adapted to differentiate between e-health and not e-health publications. BERT is a general-purpose language representation model trained in an unsupervised manner. Unlike simpler Natural Language Processing (NLP) models based on word counts, BERT takes context of tokens into consideration. This way, the model captures language semantics and is able to differentiate between tokens that would be considered the same by word-count models.

To learn contextual relationships between words (or text tokens), BERT utilizes a Transformer neural network architecture and an attention mechanism. BERT is trained using two strategies: masked word prediction in a sequence of tokens and next sentence prediction. Both strategies are trained together minimizing combines loss function. BERT model, trained on a big language corpus, learns language representation and relationships between text tokens. Such general-purpose model can be later finetuned for specific NLP tasks without the need to retrain entire model [8]. In comparison to model pretraining, finetuning is considerably less demanding computationally and can be performed on much smaller datasets.

To adapt the general purpose BERT-based model to a specific task, it was finetuned on a manually labelled publication corpus presented earlier (e-health publication dataset) [7]. This corpus contains 1891 publications (816 e-health, 1075 not e-health) papers. Text from title, keywords and abstract fields was used for model finetuning. Finetuning was performed in Google Colab environment. The e-health publication dataset was split into training (60%), validation (20%) and testing (20%). The model was evaluated using a random test set (20% of the e-health publication dataset) after model finetuning.

Due to manual data collection, the e-health publication dataset was skewed and contained a much larger proportion of e-health publications than data available in PubMed [7]. To ensure that the model generalizes for data available in PubMed, an additional evaluation step was included. A PubMed dataset, containing 924 publications (25 e-health and 899 not-e-health) was manually labelled and used for validating the model's performance. The PubMed dataset represents e-health and not e-health class distribution in data extracted from PubMed.

To map e-health publications to the focus area of the National E-health Strategy, a machine learning model developed previously was employed [7]. This model is based on token count values (Term Frequency – Inverse Document Frequency, TF-IDF) and was trained on the e-health publication dataset to differentiate between the following publication classes:

1) Digitization of work processes.
2) Seamless/coherent patient pathways.
3) Improved use of health data.
4) New ways to deliver healthcare.
5) Common foundation for digital services.
6) National e-health management and increased implementation.

III. Results

The architecture and validation of the developed models are summarized in this section.

A. System design

To support up-to-date monitoring of e-health publications, a data analytics pipeline was set up. The pipeline contains three major components (Figure 1):

1. Data import. This component handles queries to the data providers (PubMed in the current setup) and returns metadata for every publication meeting inclusion criteria. When monitoring publications from Norway, inclusion criteria was limited to at least one coauthor affiliated with Norway and publication dates (01-01-2010 – 01-04-2020). These data are stored in a relational database for further analysis.
2. The data analytics component hosts pretrained machine learning models for classifying publications into specific groups. The two-class classifier performs dataset denoising, discarding irrelevant (not e-health) publications. Since our focus is monitoring production e-health related papers, only these manuscripts are considered in further analysis. The six-class classifier classifies e-health publications into 6 focus areas of the Norwegian E-health Strategy [3].
3. The visualization module performs data aggregations and presents the results of the data analytics step in a visual way.

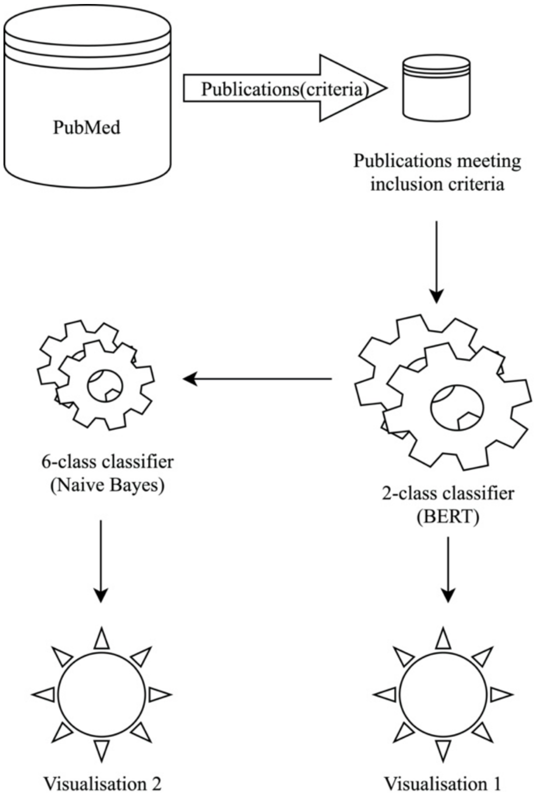
PubMed
Publications(criteria)
Publications meeting inclusion criteria
6-class classifier (Naive Bayes)
2-class classifier (BERT)
Visualisation 2
Visualisation 1

Figure 1. Data analytics pipeline

B. Performance of the classification models

The performance of the 2-class model was tested on two datasets containing different ratios of positive class (e-health) publications. The e-health publication dataset contains the same ratio of positive and negative class examples as the dataset used for model finetuning. The PubMed dataset represents a realistic ratio of positive and negative class examples observed in PubMed (Table I).

TABLE I. PERFORMANCE OF THE 2-CLASS MODEL

Dataset	Class	Precision	Recall	f-1 score	AUC
E-health publication dataset	Not e-health	0.92	0.88	0.9	0.888
	E-health	0.85	0.90	0.87	
PubMed dataset	Not e-health	0.99	0.99	0.99	0.858
	E-health	0.82	0.72	0.77	

A trained 6-class classifier, reported in an earlier publication, was used for classifying e-health publications into the focus areas of the Norwegian E-health Strategy [7]. The performance of this model is available in Table II.

TABLE II. PERFORMANCE OF THE 6-CLASS CLASSIFIER [7]

Class	Precision	Recall	f1-score
1. Digitization of work processes	0.70	0.58	0.63
2. Better continuity of care	0.61	0.62	0.62
3. Improved use of health data	0.62	0.71	0.67
4. New methods to provide healthcare	0.74	0.77	0.75
5. Common foundation for digital services	0.53	0.62	0.57
6. National e-health management and increased implementation	0.66	0.64	0.65

C. Visualizations

To illustrate how the trained model could be used for monitoring scientific publications in e-health, all publications containing "Norway" in author affiliation and published during the last 10 years (01-01-2010 – 01-04-2020) were extracted using PubMed API. More than 70 000 publications were published in PubMed by authors affiliated to Norway. The aforementioned 2-class model was used to filter out irrelevant papers based on their title, keywords, and abstract. The number of e-health publications stratified yearly are visualized in Figure 2.

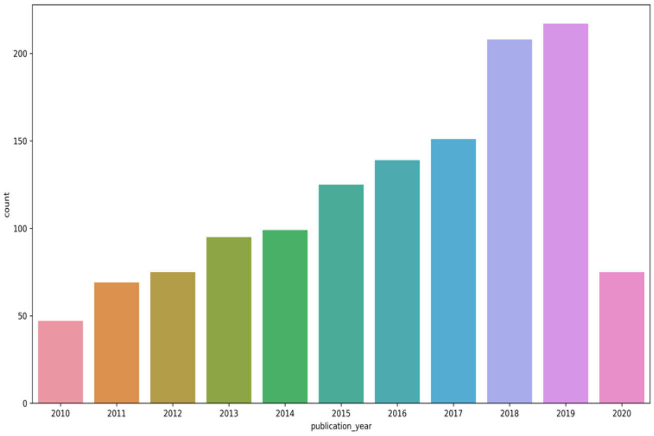

Figure 2. E-health publications authored by researchers affiliated to Norway. Publication period 01-01-2010 – 01-04-2020

To map the identified e-health publications to the focus areas of the National E-health Strategy, they were classified using a 6-class classifier (Figure 3, Figure 4) [3].

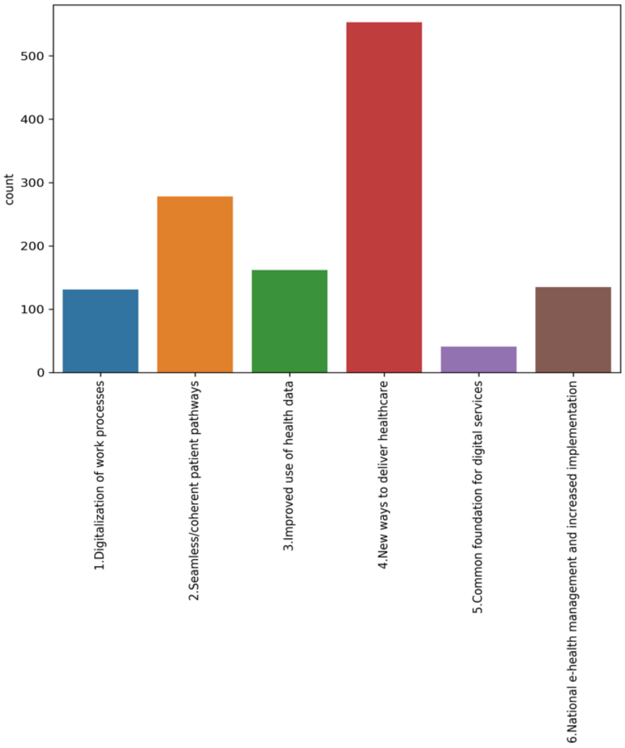

Figure 3. Classification of e-health publications into focus areas of a National E-health Strategy

To show how number of publications in each class developed in time, they were stratified on yearly basis and visualized in Figure 4.

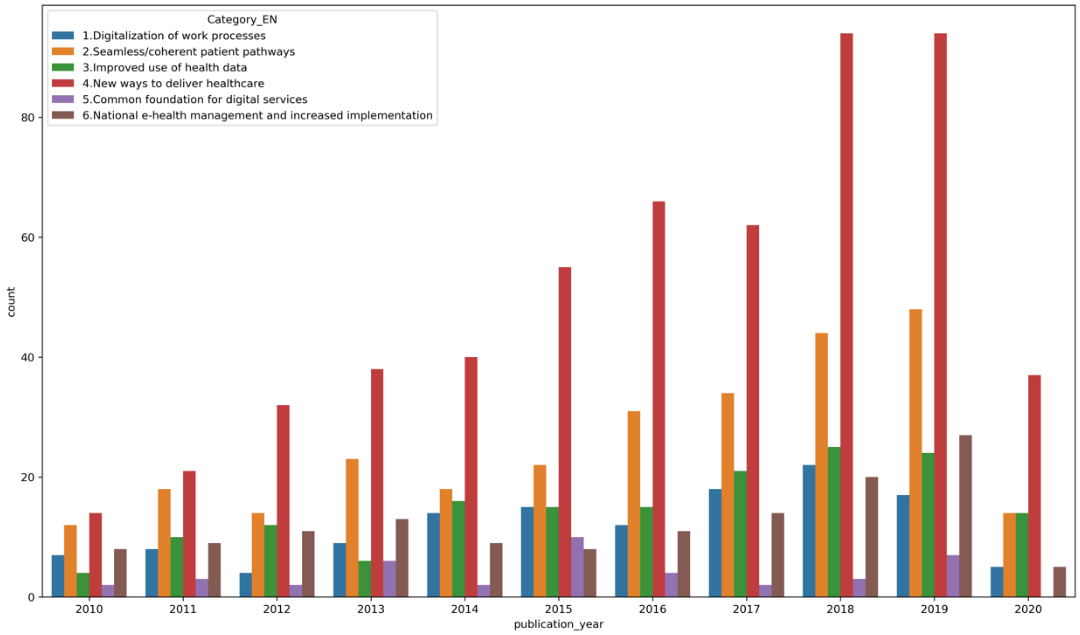

Figure 4. Classification of e-health publications into focus areas of a National E-health Strategy stratified yearly

IV. Discussion

In this paper, we demonstrated how to leverage a data analytics pipeline for monitoring the production of scientific publications. This pipeline is flexible and easily extendable to other fields, data sources and visualizations.

A. *Key findings*

Monitoring of e-health publications shows an increasing pace of publishing scientific contributions in the field of e-health in Norway (Figure 2). It may not be surprising, considering similar trends are observable in other fields of research. However, it is an important result, demonstrating the feasibility of the presented monitoring approach. We previously performed a similar experiment manually querying selected research databases with a set of keywords specifically combined to capture e-health publications [7]. Considering that manual search included more data sources, it is natural that some differences in yearly publication counts were observed when comparing these methods. Regardless of these deviations, both methods captured the same publication trend.

B. *Model performance*

Classification of e-health and not e-health publications has been attempted previously using more traditional approaches [7]. These data processing pipelines consist of two major steps: transformation of free text into numeric representations (for instance, TF-IDF) that are later used for training machine learning models. This approach is rather simple computationally, however, does not take language semantics into account. Advanced NLP models, such as BERT can address language specifics much better, however, are computationally intensive and require more data to deliver satisfactory performance. Even though pretrained BERT models are less dependent on the amount of data (they have learned language representation during pretraining), generating sufficient amount of data for model finetuning could be problematic in some fields. In our case, finetuning BERT for 2-class classification showed performance increase with regards to the baseline model based on TF-IDF [7]. However, finetuning BERT for the 6-class classification resulted in poor performance, indicating data insufficiency. The traditional NLP model based on TF-IDF features and Naive Bayes classifier performed better for this task.

C. *Limitations*

Accuracy of the machine learning model has to be taken into consideration when looking at the absolute numbers of e-health publications. Even though the model generalizes reasonably well, some performance decline is observable in the PubMed dataset (Table 1). It is caused by the different class label distribution in the e-health publication dataset used for model finetuning and PubMed dataset used for validation. While the absolute numbers presented in Figure 1 should be interpreted with caution, the identified publishing trend is not affected by these discrepancies.

Due to the difficulties of accessing other research publication databases, only papers indexed in PubMed were included in this study. Unfortunately, PubMed lacks publications indexed elsewhere or only published in non-indexed conference proceedings. However, it is fair to assume that the major part of e-health related publications is available in PubMed. E-health is often considered as an intersection of health, technology and social sciences. Focus on health topics makes PubMed the preferred database for e-health publications.

D. *Future work*

In this proof-of-concept phase, automatic data collection was implemented using PubMed API for querying relevant publication metadata. Even though the number of publications in PubMed shows general trends in scientific paper production, this number could be considered misleading in terms of absolute publication count. Previous research shows that some e-health publications are published in channels that are not indexed by PubMed [7]. Other major research databases were not included in data collection due their limitations and costs associated with consumption of metadata APIs. Inclusion of publication data from Scopus and Web of Science databases is planned for the future.

Publications from other countries could contribute to the insights delivered by this system by contrasting global e-health research and development trends. Even though the 6-class model is optimized for the Norwegian context, it could be useful for other countries. Replacing the 6-class model with another one, addressing specific use case requirements better (for instance, focus areas of e-health implementation strategy in other countries) is straightforward.

V. Conclusion

National strategies are complex and often generic documents that are difficult to map to the scientific development in a field. Academic publishing could be used as a proxy hinting to the maturity of a specific field and a direction it may take in the future. In this paper, we presented how novel data science methods could be leveraged to map production of research papers into focus areas of the Norwegian E-health Strategy. This mapping shows the coverage of various focus areas in the e-health strategy by scientific publications and provides insights to the decision makers about underresearched topics.

References

[1] WHO Global Observatory for eHealth and World Health Organization, Global diffusion of eHealth: making universal health coverage achievable : report of the third global survey on eHealth. 2016.

[2] Direktoratet for E-helse, “Utredning av «Én innbygger – én journal»,” [Report "One Citizen - One Health Record"] Dec. 2015. [Online]. Available from: http://www.webcitation.org/73nstSwCb. [Accessed: 09-Nov-2018].

[3] Direktoratet for E-helse, “Nasjonal e-helsestrategi 2017-2022,” [National e-health strategy 2017-2022] [Online]. Available from: http://www.webcitation.org/73ntEa1LR. [Accessed: 09-Nov- 2018]

[4] Direktoratet for E-helse, “Nasjonal handlingsplan for e-helse 2017-2022,” [National action plan for e-health 2017- 2022] [Online]. Available from: http://www.webcitation.org/73ntLCceo. [Accessed: 09-Nov-2018].

[5] G. Eysenbach, “What is e-health?,” J. Med. Internet Res., vol. 3, no. 2, p. e20, 2001, doi: 10.2196/jmir.3.2.e20.

[6] C. W. Belter, “Bibliometric indicators: opportunities and limits,” J. Med. Libr. Assoc. JMLA, vol. 103, no. 4, pp. 219–221, Oct. 2015, doi: 10.3163/1536-5050.103.4.014.

[7] A. Budrionis, K. Lind, I. M. Holm, O. Saadatfard, and R. Pedersen, “Establishing Baseline in the Status of E-health Research in Norway,” presented at the eTELEMED 2019, The Eleventh International Conference on eHealth, Telemedicine, and Social Medicine, Feb. 2019, pp. 85–89, Accessed: Mar. 01, 2019. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=etelemed_2019_5_40_40070.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” ArXiv181004805 Cs, May 2019, Accessed: Apr. 28, 2020. [Online]. Available: http://arxiv.org/abs/1810.04805.

ISBN: 978-1-61208-763-4