The Influence of NegEx on ICD-10 Code Prediction in Swedish: How is the Performance of BERT and SVM Models Affected by Negations?

Andrius Budrionis[1,2], Taridzo Chomutare[1], Therese Olsen Svenning[1] and Hercules Dalianis[1,3]
[1]Norwegian Centre for E-health Research, Tromsø, Norway Andrius.Budrionis@ehealthresearch.no
[2]Faculty of Science and Technology, UiT The Arctic University of Norway, Tromsø, Norway
[3]Department of Computer and Systems Sciences (DSV), Stockholm University, Kista, Sweden

Abstract
Clinical text contains many negated concepts since the physician excludes irrelevant symptoms when reasoning and concluding about the diagnosis. This study investigates the machine interpretation of negated symptoms and diagnoses using a rule-based negation detector and its influence on downstream text classification task. The study focuses on the effect of negated concepts and NegEx preprocessing on classifier performance for predicting ICD-10 gastro surgical codes assigned to discharge summaries. Based on the experiments, NegEx preprocessing resulted in a slight performance improvement for traditional machine learning model (SVM) and had no effect on the performance of the deep learning model KB/BERT.

Keywords
Clinical text, negation, NegEx, Swedish, BERT, ICD-10 diagnosis codes

1 INTRODUCTION

Physician's reasoning to find the correct diagnosis of a patient are often trying to exclude symptoms until the hopefully correct diagnosis is concluded. This leads to the patient's record containing many negations excluding various clinical concepts [1].

Negated terms make it difficult for machines to interpret natural language in patient records. One of the first approaches to identify negated symptoms and diagnoses was the development of NegEx, [2]. This relatively simple rule-based algorithm showed acceptable performance in identifying negated clinical symptoms and diagnosis. The output of NegEx (negation tags) could be an important additional feature for various Natural Language Processing (NLP) tasks.

Numerous approaches for detecting negations in free text have been proposed. In addition to the aforementioned NegEx [2], more advanced models were developed taking advantage of language semantics [3] [4]. Ettinger [5] performed fine-tuning of BERT on manually annotated data sets containing both negations and non-negations. The findings showed that BERT was not able to distinguish the negations in the text within acceptable accuracy. On contrary, Lin et al. found that fine-tuning BERT on clinical text that contains annotated negations led to BERT learning to predict negations [6].

The influence of negated symptoms and diagnosis existing in the text and the output of negation detectors on the downstream modelling tasks has only be studied to a limited extent [7] [8] [9]. It is unclear how NegEx preprocessing contributes to the overall performance of text classification models and what use cases or machine learning models benefit from negation tagging or removal from the text.

Remmer et al. [10] carried out classification experiments for predicting groups of ICD-10 diagnosis codes assigned to Swedish discharge summaries. The researchers used both traditional machine learning methods and the deep learning BERT model. The BERT model outperformed the traditional models. Since clinical text contains many negations, specifically for Swedish clinical text 13.5% of the sentences or expressions were negated [11], it would be valuable to know how these affect the classification results and also if there are methods to cope with negations. This paper studies the effect of NegEx on ICD-10 code prediction task.

2 RELATED RESEARCH

Existing research on the effect of NegEx on the downstream modelling tasks is fragmented and limited to a few publications, mostly focusing on traditional machine learning algorithms, such as Support Vector Machines (SVM) used for sentiment analysis. Sharif et al. reported significant increase in accuracy, precision and recall predicting sentiments in customer reviews after text was preprocessed by a negation detector [7]. Similar findings were reported for Naïve Bayes, Artificial Neural Network (ANN), and Recurrent Neural Network (RNN) models used for sentiment analysis. The largest positive effect of negation tagging was observed in RNN models [8]. Kaddoura et al. demonstrated that treating negations in Facebook posts resulted in 20% increase in F1-score in sentiment analysis [9]. Considering the number and importance of negations in medical narrative, similar improvements in clinical NLP tasks could be expected.

The 18th Scandinavian Conference on Health informatics, Tromsø, Norway, August 22-24, 2022. Organized by UiT The Arctic University of Norway. Conference Proceedings published by Linköping University Electronic Press at https://doi.org/10.3384/ecp187.

3 DATA AND METHODS

3.1 Data

The Stockholm EPR Gastro ICD-10 Pseudo Corpus was used in the experiments. It contains discharge summaries and their manually assigned ICD-10 gastro related diagnosis codes. Additional details on the dataset can be found in [10]. The specific corpus variant used in this paper is called Pseudo Corpus since it has been de-identified with regard to Protected Health Information, PHI, and the identified PHIs have been replaced with realistic pseudonyms [12].

The deidentification and pseudonymisation system, called HB-Deid, was used for cleaning the text from sensitive details. It detects the following PHI classes: First Name, Last Name, Age, Location, Health Care Unit, Date, Phone Number, Organisation and Social Security Number and replaces them with realistic pseudonyms or surrogates. HB-Deid is based on Conditional Random Fields algorithm and rule-based preprocessing step to find missed phone numbers and social security numbers through regular expressions. The final step in HB-Deid is the Pseudonymiser that replaces the identified entities with realistic surrogates. After deidentification the Stockholm EPR Gastro ICD-10 Pseudo Corpus can be shared with academic community. This research has been approved by the Swedish Ethical Review Authority under permission no 2021-03758.

The dataset consists of 6,002 discharge summaries from 4,985 unique patients, 813,154 tokens in total. 263 distinct class labels (ICD-10 codes) are present in the text.

3.2 ICD-10 blocks

Many classes (ICD-10 codes) were represented by very few examples. To make modelling task easier, the label space was condensed into ICD-10 code blocks combining multiple labels into a single class (K00-K14, K20-K31, K35-K38, K40-K46, K50-K52, K55-K64, K65-K67, K70-K77, K80-K87, K90-K93).

The blocks are logical partitions of the gastrointestinal domain, starting from the oral cavity to the rest of the digestive system, and are well-recognised in the medical field. Classifying notes into the blocks can be useful in clinical practice, as well as part of a pipeline to classify into more granular ICD-10 codes.

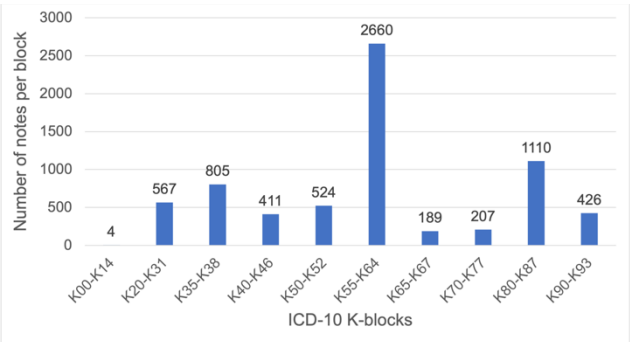

Figure 1. The distribution of number of discharge summaries per ICD-block.

The distribution of discharge summaries per block are shown in Figure 1. The majority of discharge summaries contains only one code block label; the maximum number of code block labels per discharge summary is four.

3.3 Methods

To study the effect of negations in clinical text on the performance of ICD-10 code block classification using different classifiers, the clinical notes were preprocessed using a Swedish negation detector (NegEx). The Swedish NegEx has a performance of 75.2% precision and a recall of 81.9% applied on Swedish clinical text [13].

Two versions of NegEx outputs were used in downstream modelling tasks: negated symptoms and diagnoses were either tagged or removed (referred to as tagged and removed negations in the reminder of this paper). The original (without NegEx preprocessing) dataset was used in baseline experiments.

Two types of classifiers, representing traditional and state-of-the-art machine learning models used in NLP were selected. For the traditional models, a SVM classifier (scikitlearn v1.0.2) was trained using TF-IDF vectors from clinical notes. 10-fold cross-validation was used when training the model; 10% of data was held out from training and used for testing purposes only. Data was split at a record level.

For the state-of-the-art model, a KB/BERT, a Swedish general language model pretrained on newspapers, Swedish Wikipedia and government documents, was used [14]. KB/BERT was finetuned on the preprocesssed datasets (tagged and removed negations) and tested on the holdout dataset. The parameters used with the pytorch model are shown in Table 1.

Parameter	Value
Test size	0.2
K-fold	10
Random state	42
Epochs	15
Batch size train	6
Batch size test	6
Gradient accumulation	8
Learning rate	3e-5
Warm up	400
Threshold	0.3

Table 1. The parameters used to fine tune KB/BERT model.

4 RESULTS

Experiment results are summarised in Table 2 and show some interesting trends. Tagging and removal of negated symptoms and diagnoses resulted in a small performance boost for the SVM model in comparison to the baseline. While this improvement in performance was minor, it may be sufficient for considering NegEx preprocessing as a technique for training an optimal model.

SVM code blocks

	Precision	Recall	F_1-score
Original dataset	0.89	0.61	0.73
Negations tagged	0.90	0.63	0.74
Negations removed	0.90	0.63	0.74

KB/BERT code blocks

	Precision	Recall	F_1-score
Original dataset	0.80	0.81	0.81
Negations tagged	0.81	0.80	0.81
Negations removed	0.80	0.81	0.81

Table 2. Model performance on holdout dataset. (In micro average precision, recall and F1-score respectively).

Performance of KB/BERT did not follow the same trend. Tagging and removing negations in the fine-tuning dataset did not affect model performance in terms micro average of F_1-score. Per class performance is presented in Table 3.

4.1 NegEx post-processing challenges

So far we have discussed pre-processing of data for training purposes. Another option is to use negation in post-processing, which has the potential to yield improved performance. One approach to post-processing involves analysing the negated concepts to determine the likely associated ICD-code or block. For instance, one of the most frequently negated terms is "abscess" (see Table 4), which is associated with two blocks within the Gastrointestinal ICD-10 space, namely, ['K00-K14', 'K55-K64']. Comparing predicted results against these known ICD-10 associations can help exclude possibly wrong predictions. However, in this case, there are multiple factors that presented significant challenges for post-processing, and in this subsection we discuss the top four factors.

The first challenge is related to the ICD-10 level or hierarchy, where the ICD-10 block is a high-level reference, as opposed to a lower level 4-char code such as K56.7 or simply K567. Whereas it is possible to exclude lower-level codes if the associated concept is negated, it is not logical to exclude blocks, since a block contains multiple related ICD-10 codes.

Term	Similar ICD-10 concept	Associated block(s)	Frequency
abscess	abscess [abscess]	'K00-K14', 'K55-K64'	9
kolecystit	kolecystit [cholecystitis]	'K80-K87'	7
peritonit	peritonit [peritonitis]	'K65-K67', 'K35-K38'	6
kräkning	Kräkningar [Vomiting]	'K90-K93'	4
divertikulit	Divertikel [diverticulum]	'K20-K31', 'K55-K64', 'K35-K38'	3

Table 4. Test partition's top 5 negated terms and their associated ICD-10 concepts and blocks.

The second factor is related to where the negated terms appear in the wider context that includes non-gastrointestinal blocks. Since our label-space is focused on Gastrointestinal blocks, some negated concepts will fall outside the relevant vocabulary or scope. For instance, "leukocytos" [leukocytes] is associated more with D72--"Other disorders of white blood cells", rather than gastro K-codes. Therefore, negation of this term likely will not affect the performance for the limited gastro case. However, we found that these out-of-vocabulary negations were comparatively small, making up approximately 10% all of negations in the test set.

Block	ICD-10 description	F_1-score original data		F_1-score tagged negations		F_1-score removed negations	
		SVM	BERT	SVM	BERT	SVM	BERT
K00-K14	Oral cavity, salivary glands and jaws	0.00	0.00	0.00	0.00	0.00	0.00
K20-K31	Esophagus, stomach and duodenum	0.32	0.68	0.56	0.70	0.42	0.68
K35-K38	Disease of appendix	0.91	0.95	0.93	0.95	0.92	0.95
K40-K46	Hernia	0.39	0.70	0.36	0.70	0.39	0.72
K50-K52	Non-infectious inflame. Intestine	0.32	0.72	0.37	0.73	0.26	0.73
K55-K64	Other diseases of the intestine	0.85	0.87	0.86	0.86	0.87	0.86
K65-K67	Disease of the peritoneum	0.00	0.12	0.00	0.20	0.00	0.15
K70-K77	Diseases of the liver	0.18	0.49	0.00	0.49	0.00	0.52
K80-K87	Gallbladder, bile ducts and pancreas	0.83	0.92	0.85	0.92	0.81	0.91
K90-K93	Other diseases of the digestive system	0.36	0.54	0.24	0.52	0.38	0.53
Micro avg		0.73	0.81	0.74	0.81	0.74	0.81

Table 3. Model performance on held out dataset using SVM and KB BERT using 10-fold cross validation. The summary of all classes is in micro F1-score.

The third factor is somewhat related to the second factor in that there were only a few records with negations, compared to the number of records in the test set. There were only 46 records with a total of 52 negated terms, representing approximately 7.7% of the test set. This factor also helps partially explain the generally muted performance gains after tagging negations.

The final factor relates to the performance of NegEx on the Swedish clinical text. Since NegEx does not understand complex semantic meaning of sentences, some negated terms will actually be a positive diagnosis. NegEx performance on Swedish data presents peculiar challenges that reduces the effect of any post processing. We discuss concrete examples of NegEx failures in the next subsection.

4.2 Error analysis

Here follows an error analysis on the performance of the rule-based NegEx on some of the clinical text.

First a correct negation tagging is shown below, where NegEx marks up “leukocytes” as a negated concept.

“Pat inlägges fastande. Labmässigt noterar man CRP 47, som sjunker till 8. Ingen <NEGATED>leukocytes</NEGATED>”.

(In Eng.) “Pat is admitted fasting. CRP of 47 was noted in the lab, which drops to 8. No <NEGATED> leukocytos</NEGATED>”.

Linking words such as 'but' are prone to misinterpretation. 'Not something but something else' should confirm the last condition, not exclude it as shown in the example below.

“genomgår gastroskopi som inte visar någon <NEGATED> främmande kropp </NEGATED>, däremot en <NEGATED> esofagit grad 3 </NEGATED>”.

(In Eng.) “undergoes gastroscopy which shows no <NEGATED>foreign body </NEGATED>, but a <NEGATED> esophagitis grade 3 </NEGATED>”.

Double trigger words may also lead to wrong negations. In the following example two trigger words which should null each-other out, meaning that there is in fact a condition present, gets interpreted as the opposite.

“man med smärta i buken till vänster sedan en vecka, där man inte kan utesluta <NEGATED> divertikulit </NEGATED>”.

(In Eng.) “man with abdominal pain to the left for a week, where one cannot rule out <NEGATED> diverticulitis </NEGATED>“.

5 DISCUSSION

Our findings indicate that the presence of negated symptoms and diagnoses in clinical text may have varying effect on the performance of ICD-10 code prediction tasks. While traditional machine learning model (SVM) trained using bag-of-words vector representations of clinical text experienced minor benefits of NegEx preprocessing (tagging and removal), the performance of the state-of-the-art model (KB/BERT) was not affected.

These findings may be explained by the way text is represented for the classification algorithms. A bag-of-words representation is not capable of capturing any semantic relationships between text tokens, and NegEx preprocessing enables it to differentiate between negated and non-negated concepts. Removal of negated symptoms and diagnoses discards some "noise" in the training data resulting in a small increase in classifier performance.

KB/BERT uses an underlying Swedish language model learned from a large amount of text. This language model is used when tokenizing the text and transforming tokens into embeddings capturing various language properties. These capabilities help KB/BERT differentiate between negated and non-negated clinical terms and use negations as additional features when making predictions. Therefore, tagging and removing negated symptoms and diagnoses has not affected classifier performance. KB/BERT managed to interpret the negations without using any preprocessing with NegEx and showed better results than the traditional machine learning methods combined with NegEx.

5.1 Limitations

The choice of classification algorithms comes as the main limitation of this paper. While both traditional and deep-learning-based models were studied in the experiments, the findings cannot be generalized for all algorithms in these classes. A great variety of classification algorithms belonging to both classes calls for extensive experiments using a well-grounded selection process taking properties of specific algorithms into account. SVM and KB/BERT were chosen due to their popularity for text classification tasks, therefore results presented in this paper should only be considered as preliminary findings calling for more research.

Rule based NegEx has a relatively high error rate and a more robust approach should be explored for preprocessing clinical text. More sophisticated implementations of negation detectors result in higher accuracy that may affect the performance of downstream modelling tasks [3] [4].

6 CONCLUSIONS

Regardless of the limitations of NegEx (see section 4.2 for more details), the performed experiments show the following trends. Model based on bag-of-words text representation benefited from NegEx proprocessing, resulting in increasing performance. On the contrary, performance gains were absent for the advanced model, capturing the semantic relationships between text tokens.

Post-processing was complicated by the non-atomic nature of the label space, where the ICD blocks contain multiple individual ICD codes, making it difficult to exclude any negated labels.

7 LANGUAGE RESOURCE AVAILABLE

The pseudonymised dataset, the Stockholm EPR Gastro ICD-10 Pseudo Corpus, is available for research for academic researchers after signing a confidentially agreement, please contact Hercules Dalianis, email: hercules@dsv.su.se.

8 REFERENCES

[1] Epstein, R. M. "How Doctors Think." *Journal of Clinical Investigation*, Vol. 117, No. 10, 2007, pp. 2738–2738. https://doi.org/10.1172/JCI33149.

[2] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." *Journal of Biomedical Informatics*, Vol. 34, No. 5, 2001, pp. 301–310. https://doi.org/10.1006/jbin.2001.1029.

[3] Montenegro, O., Pabon, O. S., and De Pinerez R., R. E. G. "A Deep Learning Approach for Negation Detection from Product Reviews Written in Spanish." Presented at the 2021 XLVII Latin American Computing Conference (CLEI), Cartago, Costa Rica, 2021.

[4] Khandelwal, A., and Britto, B. K. "Multitask Learning of Negation and Speculation Using Transformers." Presented at the Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Online, 2020.

[5] Ettinger, A. "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models." *Transactions of the Association for Computational Linguistics*, Vol. 8, 2020, pp. 34–48. https://doi.org/10.1162/tacl_a_00298.

[6] Lin, C., Bethard, S., Dligach, D., Sadeque, F., Savova, G., and Miller, T. A. "Does BERT Need Domain Adaptation for Clinical Negation Detection?" *Journal of the American Medical Informatics Association*, Vol. 27, No. 4, 2020, pp. 584–591. https://doi.org/10.1093/jamia/ocaa001.

[7] Sharif, W., Samsudin, N. A., Deris, M. M., and Naseem, R. "Effect of Negation in Sentiment Analysis." Presented at the 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 2016.

[8] Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S., and Sharma, R. "Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection." *Procedia Computer Science*, Vol. 185, 2021, pp. 370–379. https://doi.org/10.1016/j.procs.2021.05.038.

[9] Kaddoura, S., Itani, M., and Roast, C. "Analyzing the Effect of Negation in Sentiment Polarity of Facebook Dialectal Arabic Text." *Applied Sciences*, Vol. 11, No. 11, 2021, p. 4768. https://doi.org/10.3390/app11114768.

[10] Remmer, S., Lamproudis, A., and Dalianis, H. "Multi-Label Diagnosis Classification of Swedish Discharge Summaries – ICD-10 Code Assignment Using KB-BERT." Presented at the The International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 2021.

[11] Dalianis, H., and Velupillai, S. "How Certain Are Clinical Assessments? Annotating Swedish Clinical Text for (Un)Certainties, Speculations and Negations." Presented at the the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 2010.

[12] Berg, H., and Dalianis, H. "Augmenting a De-Identification System for Swedish Clinical Text Using Open Resources and Deep Learning." Presented at the Workshop on NLP and Pseudonymisation, Turku, Finland, 2019.

[13] Skeppstedt, M. "Negation Detection in Swedish Clinical Text: An Adaption of NegEx to Swedish." *Journal of Biomedical Semantics*, Vol. 2 Suppl 3, 2011, p. S3. https://doi.org/10.1186/2041-1480-2-S3-S3.

[14] Malmsten, M., Börjeson, L., and Haffenden, C. "Playing with Words at the National Library of Sweden -- Making a Swedish BERT." *arXiv:2007.01658 [cs]*, 2020.

9 ACKNOWLEDGEMENT

This work was partially funded by the ClinCode project at the Norwegian Centre for E-health Research. Grant number 318098, Research Council of Norway.