Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210861

Semiautomatic Identification of Pulmonary Embolism in Electronic Health Records Through Sentence Labeling

Gleb DANILOV[a,1], Timur ISHANKULOV[a], Alexandra KOSYRKOVA[a], Maria SHULTS[a], Semen MELCHENKO[a], Tatyana TSUKANOVA[a], Michael SHIFRIN[a] and Alexander POTAPOV[a]

[a]*Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*

Abstract. In this study, we tested the quality of the information extraction algorithm proposed by our group to detect pulmonary embolism (PE) in medical cases through sentence labeling. Having shown a comparable result (F1 = 0.921) to the best machine learning method (random forest, F1 = 0.937), our approach proved not to miss the information of interest. Scoping the number of texts under review down to distinct sentences and introducing labeling rules contributes to the efficiency and quality of information extraction by experts and makes the challenging tasks of labeling large textual datasets solvable.

Keywords. Natural Language Processing, Pulmonary Embolism, Neurosurgery, Machine Learning

1. Introduction

The reliability of classification or regression by machine learning based on natural language processing (NLP) largely depends on the primary textual data quality and the target variable's validity. Data labeling in medicine is usually performed by medical experts. The primary extraction of information from a body of unstructured narrative texts is a non-trivial task, especially for non-professionals in NLP [1]. Extracting information from tens of thousands of case records, even by a group of experts, can take an unreasonably enormous amount of time, and the reliability of the final result is still questionable. Our previous study demonstrated a moderate level of agreement between experts assessing the same medical records [2]. Therefore, the parallel extraction of information from one massive textual source by different healthcare professionals necessitates resolving the disagreements, which is also laborious.

Our group has previously proposed an algorithm for semi-automatic information extraction (IEA, *information extraction algorithm*) from unstructured texts written in natural language (patent RU2751993C1) [3]. The essence of the algorithm is in the decoding of distinct phrases - microcontexts (for example, n-grams) containing words

[1] Corresponding Author, Gleb Danilov, N.N. Burdenko Neurosurgery Center, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

from the lexicon, used to describe the sought phenomenon. When retrieving information about relatively rare facts, the size of the micro contexts can be expanded to complete sentences. Thus, significantly narrowing the scope of the texts to analyze down to certain sentences dramatically saves time and other costs for specific information extraction. We assume it is possible to extract the information on various phenomena related to a medical case summarizing the set of distinct sentences labeled with our algorithm. In this study, we assessed the quality of information extraction about pulmonary embolism (PE) following neurosurgical interventions using the proposed method of information extraction based on sentences labeling.

2. Methods

The data for the study were obtained from the Electronic Health Records (EHR) of the National Medical Research Center of Neurosurgery, named after academician N.N. Burdenko (Moscow, Russia) for the period between 2000 and 2017 (90 688 cases). We queried the database with keywords related to the PE lexicon in Russian. Thus, the clinical narrative texts in which PE was mentioned or suspected were retrieved. The clinical corpus was constructed with postoperative daily notes, examinations, postmortem findings, and other relevant documents typed on a keyboard with free text. All health records were screened by three similarly skilled neurosurgeons independently. The experts with an overall working experience of 7-11 years had been previously trained in neurosurgical residency at N.N. Burdenko Neurosurgery Center. These coders had small comparable experience in labeling data from approximately 2000 health records. However, no expert was specially trained for PE detection task. The texts were presented to the experts in a special software designed to focus on medical texts only. Each expert was asked to label the cases with either “PE detected” or “No PE detected” or “the fact of PE could not be well-verified.” The disagreements between the experts were resolved with the involvement of the fourth expert.

All the records were also independently screened by the first author of this article, with the methodology described in our previous publications [2–5]. The text was tokenized into sentences and then into lemmatized words. The unique lemmas were screened to select those likely related to or certainly used in the PE description. All the sentences containing the initial words matching all selected lemmas were then reviewed to label with “PE detected” or “No PE detected” or “the fact of PE could not be well-verified” categories and score by 1, 0 and 0.5, respectively. To apply our information extraction algorithm (IEA) as described previously, we labeled the whole medical case with “PE detected” if at least one sentence in a case was assigned with the “PE detected” label. We marked the case with the “the fact of PE could not be well-verified” tag if the same sentence label occurred and no "PE detection" labels were met. We rejected PE in a case if no labeled sentences were present or “No PE detected” was the only sentence label found.

We also decided to test whether it is possible to improve decision-making on PE detection in the entire medical case with labeled sentences using machine learning. To accomplish this, the scores of labeled sentences were further summarized for each clinical case into a set of 33 parameters (maximum scores for each type of clinical document, maximum score for all documents, the simultaneous 0.5 and 1 maximums in different documents of the same case). The resulting matrix of 621 rows (clinical cases) and 34 columns (aggregated numeric features and the target variable - the presence/absence/uncertainty of PE, agreed between experts) was then used to train and test machine learning algorithms to classify cases by PE labels.

A total of 6 machine learning models were tested: random forest (RF), logistic regression (LR), support vector machine (SVM) with different kernel types (linear (lin), radial (rad), and polynomial (poly)), and K-nearest neighbors (KNN). Each model was trained in 300 experiments with resampling. We kept the training/testing sampling ratio as 80% / 20% and applied stratification by the target variable. The average sensitivity (SENS, also referred to as recall), specificity (SPEC), accuracy (ACC), positive predictive value (PPV, also referred to as precision), and F-score (F1) were computed for each model across all the experiments. To fairly compare the machine learning results with our rule-based PE detection (IEA), we calculated these indicators for the entire set of cases and reproduced the average metrics for our algorithm performance on 300 testing samples simulated as 20% of the initial sample randomly with stratification.

3. Results

A set of documents for 621 medical cases was obtained from EHR for the study and reviewed by three neurosurgeons. A total of 3161 sentences containing preselected PE lexicon was labeled by the fourth expert independently. The results of the IEA application for PE detection compared to machine learning performance are demonstrated in Table 1.

Table 1. The quality metrics of the rule-based algorithm (IEA) proposed to detect PE in medical records by sentence labeling compared to machine learning over labeled sentences.

Model	SENS	SPEC	PPV	ACC	F1
IEA (full set)	0.966	0.974	0.891	0.936	0.921
IEA (resampling)	0.967	0.975	0.893	0.937	0.921
RF	0.959	0.976	0.920	0.950	0.937
SVM (lin)	0.960	0.974	0.920	0.949	0.937
SVM (poly)	0.917	0.959	0.909	0.934	0.911
SVM (rbf)	0.739	0.903	0.859	0.873	0.759
LR	0.720	0.897	0.806	0.864	0.729
KNN	0.602	0.838	0.726	0.797	0.624

A confusion matrix for PE detection with the proposed IEA in the entire set of cases is shown in Figure 1.

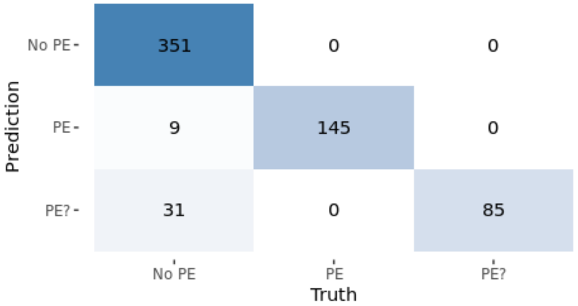

Figure 1. A confusion matrix to assess PE detection quality with the proposed IEA. “PE” = “PE detected”; “No PE” = “No PE detected”; “PE?” = “the fact of PE could not be well-verified”.

4. Discussion

We found the PE lexicon in only 621 cases out of 90688, proving PE rarity in neurosurgery (<1%) [6]. RF and SVM (lin) methods slightly outperformed our algorithm in metrics (F1 = 0.937 vs. 0.921). However, as one can notice from Fig. 1, no false rejections on PE were met, which means the algorithm is highly specific not to miss any case of interest. This fundamental property becomes especially important when searching for rare events through tens and hundreds of thousands of cases, the overwhelming majority of which does not contain the required information a priori. Moreover, that minor part of cases falling under the "suspicion" of the algorithm can be verified by a team of experts in a reasonable time. Our method demonstrates high quality considering the results of other authors [7,8]. This approach enables reliable information extraction from thousands of case histories and shaping trustworthy target variables, significantly reducing time costs. For example, assuming that a set of 621 medical cases we studied contain most PEs detected in 2000 - 2017 at N.N. Burdenko Neurosurgery Center, applying IEA to the entire database, enables labeling all the 90 688 cases at a little added time cost, which is a significant achievement.

5. Conclusions

Scoping the number of texts under review down to distinct sentences and introducing labeling rules contributes to the efficiency and quality of information extraction by experts and makes the challenging tasks of labeling large textual datasets solvable. *This project was supported by the RFBR grant 18-29-22085.*

References

[1] Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med. Informatics 2020; 8:. doi:10.2196/17984.

[2] Danilov G, Kosyrkova A, Shults M, Melchenko S, Tsukanova T, Shifrin M, Potapov A. Inter-Rater Reliability of Unstructured Text Labeling: Artificially vs. Naturally Intelligent Approaches. Stud. Health Technol. Inform. 2021;281:118–122. doi:10.3233/SHTI210132.

[3] Danilov G, Shifrin M, Strunina U, Pronkina T, Potapov A. An Information Extraction Algorithm for Detecting Adverse Events in Neurosurgery Using Documents Written in a Natural Rich-in-Morphology Language. Stud. Health Technol. Inform. 2019; 262: 194–197. doi:10.3233/SHTI190051.

[4] Danilov G, Shifrin M, Strunina Y, Kotik K, Tsukanova T, Pronkina T, Ishankulov T, Makashova E, Kosyrkova A, Melchenko S, Zagidullin T, Potapov A. Semiautomated approach for muscle weakness detection in clinical texts, in: Stud. Health Technol. Inform., IOS Press, 2020;55–58. doi:10.3233/SHTI200492.

[5] Danilov G, Shifrin M, Strunina Y, Kotik K, Tsukanova T, Pronkina T, Ishankulov T, Makashova E, Kosyrkova A, Potapov A. Detection of muscle weakness in medical texts using natural language processing, in: Stud. Health Technol. Inform., IOS Press, 2020;163–167. doi:10.3233/SHTI200143.

[6] Khan NR, Patel PG, Sharpe JP, Lee SL, Sorenson J. Chemical venous thromboembolism prophylaxis in neurosurgical patients: An updated systematic review and meta-analysis. J. Neurosurg. 2018;129:906–915. doi:10.3171/2017.2.JNS162040.

[7] Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger M-H. Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int. J. Med. Inform. 2018;117:96–102. doi:10.1016/j.ijmedinf.2018.06.002.

[8] Shen F, Larson DW, Naessens JM, Habermann EB, Liu H, Sohn S. Detection of Surgical Site Infection Utilizing Automated Feature Generation in Clinical Notes. J. Healthc. Informatics Res. 2019;3:267–282. doi:10.1007/s41666-018-0042-9.