Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210132

Inter-Rater Reliability of Unstructured Text Labeling: Artificially vs. Naturally Intelligent Approaches

Gleb DANILOV[a,1], 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. Unstructured medical text labeling technologies are expected to be highly demanded since the interest in artificial intelligence and natural language processing arises in the medical domain. Our study aimed to assess the agreement between experts who judged on the fact of pulmonary embolism (PE) in neurosurgical cases retrospectively based on electronic health records and assess the utility of the machine learning approach to automate this process. We observed a moderate agreement between 3 independent raters on PE detection (Light's kappa = 0.568, p = 0). Labeling sentences with the method we proposed earlier might improve the machine learning results (accuracy = 0.97, ROC AUC = 0.98) even in those cases that could not be agreed between 3 independent raters. Medical text labeling techniques might be more efficient when strict rules and semi-automated approaches are implemented. Machine learning might be a good option for unstructured text labeling when the reliability of textual data is properly addressed. *This project was supported by the RFBR grant 18-29-22085.*

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

1. Introduction

Unstructured text labeling is an essential and the most time-consuming step when preparing for its classification with machine learning in any professional area. We expect data labeling technologies that address unstructured medical texts to be highly demanded as the interest in artificial intelligence and natural language processing arises in the medical domain.

Medical records may provide a lot of valuable information within the free text. However, its evaluation might depend on the expert's individual experience and, possibly, a subjective perspective. Reliable information extraction or data labeling requires unambiguous professional interpretation of medical content by experts. Our study aimed to assess the agreement between experts who judged on the fact of pulmonary embolism

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

in neurosurgical cases retrospectively based on electronic health records and assess the utility of the machine learning approach to automate this process.

2. Methods

The dataset for the study was obtained from electronic health records (EHR) of N.N. Burdenko Neurosurgery Center limited by a period between 2000 and 2017 (90,688 cases treated). We queried the medical cases where the key string patterns in Russian referred to as «pulmonary embolism» (PE) were met in narrative text. As a result, 621 cases were identified in which the description of PE was anticipated.

The dataset consisted of clinical notes, doctors' reports, examinations, postmortem findings, and other clinical records sections typed on a keyboard with a free text in the period following the neurosurgical procedure. All the records were independently screened by 3 similarly skilled neurosurgeons. Expert selection criteria were the training in neurosurgical residency accomplished at N.N. Burdenko Neurosurgery Center and the working experience of 7-11 years. The coders had small and equal prior experience labeling data from approximately 2000 medical cases. However, they were not specially trained for PE detection task. The texts were presented to the experts in a special software designed to focus only on medical texts instead of the traditional EHR user interface. 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 latest cases were excluded from the current analysis.

The inter-rater reliability was assessed using Light's kappa coefficient for 3 coders and Cohen's kappa for 2 raters [1]. The evaluation of agreement level was made in a fully-crossed design [2]. The part of the dataset in which all the estimations were fully agreed upon between the independent coders was used to test a logistic regression model as a labeling automation method. The rest of the dataset which was not equally labeled by all three and independent raters was screened by the first author of this article with the methodology described in our previous publications [3]–[5]. The text was tokenized into sentences and then into words lemmatized with MyStem technology [6]. The unique lemmas were screened to select those likely related to or certainly used in the PE description. A corresponding set of sentences containing the initial words matching all lemmas selected was then reviewed to label it with "PE detected" or "No PE detected" or "the fact of PE could not be well-verified" marks. The latest were excluded from the analysis. It is important that a strict rule was implemented in this process: the "PE detected" label was applied only if the information in the sentence was stated firmly. "No PE detected" was used when no relevance of PE was discussed in a sentence. "the fact of PE could not be well-verified" was set in all cases, in which the uncertainty of the PE statement was contained in the manner of writing (e.g., accompanied by "probably", "can't be excluded", "PE?" etc.). That labeled sentence set (without uncertain cases) was used to train and test the logistic regression model to automate sentence labeling.

Both logistic regression models were trained with LASSO regularization in a loop of 300 resamples [7]. The sensitivity (SENS, also referred to as recall), specificity (SPEC), accuracy (ACC), positive predictive value (PPV, also referred to as precision), negative predicted value (NPV), F-measure (F), cutoff value (CO) and area under ROC-curve (AUC) were computed for every iteration. The bounds of 95% AUC confidence interval were calculated using bootstrapping. The cutoff point was chosen to maximize

the sum of sensitivity and specificity for every algorithm. The final metrics were averaged across 300 iterations.

All the data extracted from EHR were processed and analyzed within the R programming environment (version 4.0.3) in RStudio Server IDE (version 1.3.1093) using *irr, rsample, drlib, broom, yardstick, glmnet, doMC, dplyr, tidyr, stringr, tidytext, tidyverse, quanteda, qdapRegex, tm, scales, and cutpointr* packages.

3. Results

We observed a moderate inter-rater agreement between all three raters when labeling 621 cases (Light's kappa = 0.568, p = 0). The pairwise level of agreement between experts reached almost the same level (Cohen's kappa: 0.620 between raters 1 and 2, 0.510 between raters 2 and 3, 0.575 between raters 1 and 3, $p < 0.02$). The agreement on "PE detected" or "No PE detected" cases only was higher (n = 419, Light's kappa = 0.918, $p < 0.001$).

The experts agreed with each other fully and independently in 420 cases. The efficiency of the logistic regression model trained with regularization on that dataset (after 18 uncertain cases exclusion) is shown in Table 1. That model exploited the set of a few texts for one clinical case matched with one label (PE confirmed in 78 cases and firmly not detected in 324).

There was no equal agreement between 3 experts observed in 201 cases. Excluding 70 cases of uncertainty, splitting the text into sentences and labeling them with the method we proposed earlier led us to a dataset of 1565 sentences, with 152 indicating PE and 1413 pointing out its absence. The logistic regression trained and tested on distinct labeled sentences (as opposed to set of texts) showed much better performance in terms of all metrics except for sensitivity (see Table 1).

Table 1. The quality metrics of logistic regression models with LASSO regularization.

Method	CO	SENS	SPEC	ACC	PPV	NPV	F	ROC AUC [95% CI]
Ful agreement (n = 402)	0,17	0.859	0.894	0.888	0.667	0.965	0.744	0.936 [0.870; 0.984]
No equal agreement (n = 131)	0,46	0.745	0.991	0.967	0.905	0.973	0.813	0.980 [0.948; 0.998]

***** for all algorithms, 'PE' cases were considered when output ≥ cutoff point

Figure 1 demonstrates the superimposed ROC curves in the experiments with sets of texts equally coded by experts and with sets of sentences labeled using a special technique by additional independent expert for the cases of disagreement.

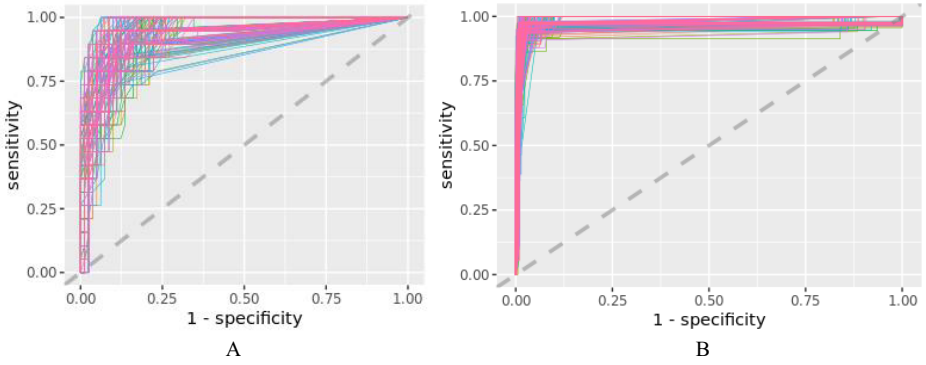

Figure 1. ROC curves superimposed after 300 times of resampling. A – logistic regression model trained and tested on a sample of 402 cases with the complete agreement between 3 experts, B – logistic regression model trained and tested on a sample of 131 cases not fully agreed between 3 experts - with sentences labeled. The improvement of performance is observed from A to B.

4. Discussion

PE is one of the most dangerous life-threatening complications in neurosurgery yet rare and possibly preventable [8]. PE research in neurosurgery requires a significant number of cases accumulated for a long period of time. Retrospective analysis of PE might be valuable but is obviously limited by the quality and completeness of medical records (non-modifiable factor) and by the qualification of experts extracting information. The results of our work demonstrate a moderate agreement (incomplete in 32,4% of cases) between 3 experts with a similar level of professional knowledge and experience. This can be partially explained by the interpretation complexity of certain cases due to a lack of information. In a number of situations, an expert can draw a conclusion based on indirect signs of PE which is discussible. It is important that no expert had been instructed on methods of information evaluation prior to the study. Our observations stress the human factor as it might contribute to the result of machine learning.

We have found that strict rules implemented specifically for a certain labeling task can make the process more efficient. Thus, special pretraining might be considered for the experts. The labeling technique we proposed earlier proved its efficiency with a traditional machine learning approach. In that case the whole sentences were labeled (instead of n-grams) which is one of possible modifications of our algorithm.

The limitations of our study are the relatively small sample sizes, different algorithms applied to different samples, sentence labeling by one expert, one type of model tested in machine learning. The state-of-art approaches (e.g., BERT) might be beneficial in future experiments. We also left the cases of uncertainty in experts' conclusions beside the scope of this article. We believe that these cases deserve a thorough consideration in a special research.

5. Conclusion

Medical text labeling techniques might be more efficient when strict rules and semi-automated approaches are implemented. Machine learning might be a good option for unstructured text labeling when the reliability of textual data is properly addressed. *This project was supported by the RFBR grant 18-29-22085.*

References

[1] Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychol. Bull. 1971 Nov;76(5):365–377.
[2] Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial A Primer on IRR. Tutor Quant Methods Psychol. 2012;8(1):23–34.
[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 Jul;262:194–197.
[4] Danilov G et al. Detection of muscle weakness in medical texts using natural language processing. Studies in Health Technology and Informatics. 2020;270:163–167.
[5] Danilov G et al. Semiautomated approach for muscle weakness detection in clinical texts. Studies in Health Technology and Informatics. 2020;272:55–58.
[6] MyStem - Yandex Technology [Online]. Available at: https://yandex.ru/dev/mystem/, Accessed 14-Jun-2020].
[7] Tibshirani R. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B. 1996;58(1):267–288.
[8] Khan NR, Patel PG, Sharpe JP, Lee SL, Sorenson J. Chemical venous thromboembolism prophylaxis in neurosurgical patients: An updated systematic review and meta-analysis. Journal of Neurosurgery. 2018 Oct;129(4):906–915.