J. Mantas et al. (Eds.)

doi:10.3233/SHTI230508

Data Quality Estimation Via Model Performance: Machine Learning as a Validation Tool

Gleb DANILOV[a,1], Konstantin KOTIK[a], Michael SHIFRIN[a], Yulia STRUNINA[a], Tatiana PRONKINA[a], Tatiana TSUKANOVA[a], Vladimir NEPOMNYASHIY[a], Nikolay KONOVALOV[a], Valeriy DANILOV[b] 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*
[b] *Kazan State Medical University, Kazan, Russian Federation*
ORCiD ID: Gleb Danilov https://orcid.org/0000-0003-1442-5993

Abstract. In our recent study, the attempt to classify neurosurgical operative reports into routinely used expert-derived classes exhibited an F-score not exceeding 0.74. This study aimed to test how improving the classifier (target variable) affected the short text classification with deep learning on real-world data. We redesigned the target variable based on three strict principles when applicable: pathology, localization, and manipulation type. The deep learning significantly improved with the best result of operative report classification into 13 classes (accuracy = 0.995, F1 = 0.990). Reasonable text classification with machine learning should be a two-way process: the model performance must be ensured by the unambiguous textual representation reflected in corresponding target variables. At the same time, the validity of human-generated codification can be inspected via machine learning.

Keywords. Neurosurgery, neurosurgical procedures, classification, machine learning, deep learning, artificial intelligence

1. Introduction

In our recent study, the attempt to classify neurosurgical operative reports into routinely used expert-derived classes exhibited an F-score not exceeding 0.74 [1]. However, the target classification could possess drawbacks, as it had been emerging for eighteen years with altering principles. In a simulation study, we showed the theoretical possibility of achieving high accuracy in operative reports classification with deep learning when the texts contained typical patterns corresponding to classes [1]. This study aimed to test how improving the classifier (target variable) affected the short text classification with deep learning on real-world data.

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

2. Methods

The operative reports with titles, bodies, and expert classes were obtained from the electronic health records of the National Medical Research Center of Neurosurgery, named after academician N.N. Burdenko (Moscow, Russia) for the period 2000 - 2017. All the characters except Cyrillic letters and single spaces were removed, and texts were tokenized with a space separator. Stop-words and tokens that occurred less than six times in the corpus were eliminated. We corrected the spelling with the method proposed in our previous work and lemmatized all word tokens [2].

All the reports were reviewed to map with a new classification (target variable) based on three strict principles when applicable: pathology type, localization, and manipulation type. After reclassification, we independently subsetted the report titles and bodies along with the old and new target variables selecting the top 13, 23, and 51 most common classes (12 datasets overall). Then the bidirectional recurrent neural networks with a gated recurrent unit (BiRNN-GRU) were trained to determine the classes separately on the titles and operative reports (12 models total).

The independent testing dataset was 25% of the initial preprocessed dataset in every machine learning setup. The rest was split into training (75%) and validation (25%) sets. The BiRNN-GRU model utilized the entire vocabulary of word tokens. It was trained with 120 epochs, batch size of 1024, 64 units in GRU layer, "softmax" activation function, categorical cross-entropy loss function, dropout and recurrent dropout rates equal to 0.4. Pre-trained FastText embeddings were applied to set weights in the embedding layer without freezing. We used word vectors that well-reflected semantic relations between terms [3]. The accuracy (ACC), sensitivity (SENS, also referred to as recall), specificity (SPEC), positive predictive value (PPV, also referred to as precision), negative predicted value (NPV), F-measure (F1), the area under ROC-curve (ROC-AUC), and area under the precision-recall curve (PR-AUC) were computed in tests.

We performed the analysis with R programming language (version 4.1.3) in RStudio Server IDE (version 2022.07.2) using *tidyverse, tidytext, textdata, tidymodels, textrecipes, furrr, catboost, tensorflow* and *keras* packages. FastText vector representations were obtained within the Python environment (version 3.6.10) in Jupyter Notebook (version 6.1.4) with fasttext library. Models were trained on the NVIDIA DGX A100 supercomputer.

3. Results

A total of 90 685 primary neurosurgical procedures were performed from 2000 to 2017. The quality metrics of deep learning classification for old (labeled as "Old" in Dataset column) and new (marked as "New" in Dataset column) target variables in samples with various numbers of classes trained over operative report titles (labeled as "Titles" in Dataset column) and bodies (marked as "Bodies" in Dataset column) independently are shown in Table 1. The quality was much better in "New" reclassified datasets compared to "Old" and more efficient for "Titles" than "Bodies." The quality tended to decrease with the number of classes growing. The best result was shown for the classification of "Titles" into 13 classes in the "New" dataset (accuracy = 0.995, F1 = 0.990). Artificial reduction in the size of that sample with stratification to make it comparable to the "Old" 13-class "Titles" sample did not influence the metrics tremendously (compare the first three rows in Table 1).

Table 1. The quality metrics for deep learning models on various subsets with different number of classes. SSize – the sample size of each subset; NC – the number of classes in each subset. * - first sample artificially reduced with stratification.

Data	SSize	NC	ACC	SENS	SPEC	NPV	PPV	F1	ROC AUC	PR AUC
Titles (New)	79 721	13	0.995	0.989	0.999	0.999	0.991	0.990	0.999	0.996
Titles (New)*	33 481	13	0.993	0.978	0.999	0.999	0.985	0.981	0.999	0.996
Titles (Old)	33 432	13	0.956	0.944	0.996	0.996	0.957	0.948	0.996	0.969
Bodies (New)	79 676	13	0.979	0.950	0.997	0.998	0.961	0.955	0.997	0.975
Bodies (Old)	33 430	13	0.954	0.948	0.996	0.996	0.948	0.947	0.997	0.972
Titles (New)	81 768	23	0.992	0.967	0.999	0.999	0.978	0.972	0.999	0.990
Titles (Old)	46 998	23	0.918	0.903	0.996	0.996	0.912	0.906	0.996	0.940
Bodies (New)	81 725	23	0.966	0.860	0.998	0.998	0.870	0.862	0.994	0.902
Bodies (Old)	47 009	23	0.889	0.866	0.995	0.995	0.869	0.865	0.994	0.910
Titles (New)	81 768	51	0.980	0.954	0.999	0.999	0.945	0.949	0.999	0.977
Titles (Old)	65 277	51	0.862	0.825	0.997	0.997	0.848	0.828	0.995	0.866
Bodies (New)	84 457	51	0.915	0.786	0.998	0.998	0.817	0.800	0.992	0.831
Bodies (Old)	65 290	51	0.823	0.770	0.996	0.997	0.808	0.788	0.994	0.820

4. Discussion

Machine learning has been rarely used for surgical procedure classification with natural language processing [1,4,5]. In our study, redesigning the target variable significantly improved the classification results and demonstrated the ability of deep learning to distinguish between human-distinguishable classes. To a certain extent, this result supports the hypothesis that we could not achieve perfect results with the previous classification because of its drawbacks. E.g., it could contain some codes overlapping by semantics, or the codes could have been misapplied due to a human factor.

The experiments we performed had inevitable limitations. The sample size was smaller in “Old” datasets than in “New” with the same number of classes. We did not augment the number of instances in “Older” classes. However, down-sampling “New” datasets almost twice did not change the quality drastically. Also, the increase in sample size (along with the number of classes) did not improve the performance, as shown in Table 1.

These results led to the conclusion: texts that are well-separated in meaning should be well-classified by machine learning. If an automated classification is poor, one should check the potential problem with the quality of texts or targets. This is how the validity of human-generated codification can be inspected via machine learning.

5. Conclusion

Unstructured text classification is a suitable task for good automation. However, reasonable text classification with machine learning should be a two-way process: the model performance must be ensured by the unambiguous textual representation reflected in corresponding target variables.

6. Acknowledgements

The machine learning experiments were supported the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-15-2021-1343.

References

[1] Danilov G, Kotik K, Shifrin M, Strunina Y, Pronkina T, Tsukanova T, Nepomnyashiy V, Konovalov N and Potapov A. Multinomial Classification of Neurosurgical Operations Using Gradient Boosting and Deep Learning Algorithms. Stud. Health Technol. Inform. 2022;295:418–421. doi:10.3233/SHTI220754.

[2] Danilov G, Shifrin M, Strunina U, Pronkina T and Potapov A. An Information Extraction Algorithm for Detecting Adverse Events in Neurosurgery Using Documents Written in a Natural Rich-in-Morphology Language, 2019. doi:10.3233/SHTI190051.

[3] Danilov G, Kotik K, Shifrin M, Strunina Y, Pronkina T, Tsukanova T, Ishankulov T, Shults M, Makashova E, Latyshev Y, Konovalov N and Potapov A. A Comparison of Word Embeddings to Study Complications in Neurosurgery, 2022. doi:10.3233/SHTI210845.

[4] Millan-Fernandez-Montes A, Perez-Rey D, Hernandez-Ibarburu G, Palchuk MB, Mueller C and Claerhout B. Mapping clinical procedures to the ICD-10-PCS: The German operation and procedure classification system use case. J. Biomed. Inform. 2020;109:103519. doi:10.1016/J.JBI.2020.103519.

[5] Khaleghi T, Murat A and Arslanturk S. A tree based approach for multi-class classification of surgical procedures using structured and unstructured data. BMC Med. Inform. Decis. Mak. 2021;21:1–12. doi:10.1186/S12911-021-01665-W/TABLES/2.