The Importance of Health Informatics in Public Health during a Pandemic
J. Mantas et al. (Eds.)

doi:10.3233/SHTI200526

The Classification of Scientific Literature for Its Topical Tracking on a Small Human-Prepared Dataset

Gleb DANILOV [a,1], Timur ISHANKULOV [a], Yuriy ORLOV [b], Mikhail SHIFRIN [a], Konstantin KOTIK [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*
[b] *Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russian Federation*

Abstract. The number of scientific publications is constantly growing to make their processing extremely time-consuming. We hypothesized that a user-defined literature tracking may be augmented by machine learning on article summaries. A specific dataset of 671 article abstracts was obtained and nineteen binary classification options using machine learning (ML) techniques on various text representations were proposed in a pilot study. 300 tests with resamples were performed for each classification option. The best classification option demonstrated AUC = 0.78 proving the concept in general and indicating a potential for solution improvement.

Keywords. Text classification, neurosurgery, machine learning, topic modeling, natural language processing, artificial intelligence

1. Introduction

The amount of scientific literature is growing rapidly each year, making it impracticable to process the entire amount of publications even within one professional domain [1], [2]. Selecting articles related to a specific subject or more stringent inclusion criteria is the most time-consuming stage when analyzing the literature, reaching up to 1-2 years in cases of systematic reviews.

The selection of literature for a certain purpose can be addressed as a binary classification task, in which one class is represented by articles of particular interest (e.g. included in a systematic review), and the second - by papers not suitable for a specific task. The article selection principle may be strictly formalized or implicitly defined, for example, related to the scientific interest of the researcher.

Our study aimed to assess the quality of solving such a task on a small real-world human-derived dataset using traditional machine learning methods as a proof of the concept.

[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 dataset was initially obtained from the PubMed search engine while performing a systematic review of artificial intelligence (AI) applications in neurosurgery in July 2019. We applied a broad definition of artificial intelligence in our search strategy: ("neurosurgical procedures" OR "neurosurgery") AND ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "NLP" OR "text mining" OR "fuzzy logic" OR "data mining" OR "big data" OR "topic model"). The publications obtained with that query were manually divided into two classes. An article was assigned to the first class (to be included in the review) in accordance with the following criteria:

- original research peer-reviewed article;
- abstract in English was available;
- the pathology/treatments discussed in the article were directly related to neurosurgery;
- the paper reported the results of AI assessment in diagnosis, treatment, prognosis, rehabilitation, or prevention.

The second class contained all other publications, which did not meet these requirements.

We proposed 19 classification options using machine learning (ML) techniques on various text representations. A document-term matrix (DTM) derived from article abstracts was a basic data structure for learning. Thus, any document was represented as a transformed DTM-vector reflecting the quantitative distribution of each term, term frequency—inverse document frequency (TF-IDF), the quantitative distribution of character N-grams, TF-IDF transformed vector of character N-grams, normalized document vectors based on character N-grams, word2vec embedding and singular value decomposition (SVD) vectors of pointwise mutual information (PMI) representation. We used character trigrams in our experiments.

In our experiments, the training subset was randomly sampled as 80% of the initial dataset (n = 671), while the remaining 20% were kept as a testing subset. The data from the training subset were transformed into several vector representations and piped into the range of ML algorithms. We applied the 5-fold cross-validation (CV) for every model on the training subset. The quality of each model was finally evaluated on a testing subset. Such an experiment was repeated three hundred times to estimate the average quality of every classification approach.

We have tested the most common ML algorithms: linear support vector machine (SVM), logistic regression and random forest. Additionally, we tested the Euclidean distance between normalized character N-gram vectors and a reference vector calculated from the training vector space as a classifier.

All the calculations were performed using the Python programming language (version 3.7) with the *pandas, numpy, ntlk, sklearn and spacy* libraries in Jupyter Notebook.

3. Results

A total of 671 articles were assigned to the first (n = 364) or second (n = 307) class. The results within each test series of 300 tests were averaged for each of 19 ML models. Averaged results are shown in Table 1. The tokenization and document vectorization

methods are referred to in the “token” and “vectorizer” columns, respectively. The mean accuracy of k-fold cross-validation (CV) on a training set is shown in the “CV” column. The accuracy (ACC), precision (PREC), recall (REC), F1-score, and the area under the receiver operating characteristic curve (AUC) obtained on the testing dataset are referred accordingly. The results in Table 1 are shown in AUC decreasing order.

Table 1. Classification results by different machine learning methods over various vector representations of tokens and documents averaged within each test series.

#	Method	Tokens	Vectorizer	CV	ACC	PREC	REC	F1	AUC
1	Linear SVC	Words	TF-IDF	0.78	0.78	0.78	0.78	0.78	0.78
2	Logistic Regression	Words	TF-IDF	0.78	0.78	0.78	0.78	0.78	0.78
3	Linear SVC	Char N-grams	TF-IDF	0.75	0.76	0.76	0.76	0.76	0.76
4	Random Forest	Words	Count	0.76	0.76	0.76	0.76	0.76	0.76
5	Logistic Regression	Words	Count	0.75	0.76	0.76	0.76	0.75	0.76
6	Logistic Regression	Char N-grams	TF-IDF	0.75	0.76	0.76	0.75	0.75	0.75
7	Random Forest	Words	TF-IDF	0.76	0.75	0.76	0.75	0.75	0.75
8	Random Forest	Char N-grams	TF-IDF	0.75	0.75	0.75	0.75	0.75	0.75
9	Linear SVC	Words	Count	0.73	0.74	0.73	0.73	0.73	0.73
10	Random Forest	Char N-grams	Count	0.74	0.74	0.74	0.73	0.73	0.73
11	Random Forest	Words	SVD/PMI	0.72	0.72	0.72	0.72	0.72	0.72
12	Logistic Regression	Words	SVD/PMI	0.68	0.72	0.71	0.71	0.71	0.71
13	Logistic Regression	Char N-grams	Count	0.70	0.70	0.70	0.70	0.70	0.70
14	Linear SVC	Words	SVD/PMI	0.67	0.69	0.69	0.69	0.69	0.69
15	Linear SVC	Char N-grams	Count	0.69	0.69	0.69	0.69	0.69	0.69
16	Random Forest	Words	word2vec	0.67	0.68	0.68	0.68	0.68	0.68
17	Linear SVC	Words	word2vec	0.56	0.56	0.52	0.53	0.43	0.53
18	Logistic Regression	Words	word2vec	0.55	0.56	0.61	0.52	0.41	0.52
19	Euclidean distance with reference	Char N-grams	Count	—	0.50	0.50	0.50	0.50	0.50

The best result (AUC = 0.78) was observed for a linear SVC algorithm trained on TF-IDF vector representations. However, logistic regression with word vectors (AUC = 0.78) weighted by TF-IDF and Linear SVC on character N-grams (AUC = 0.76) led to a close performance.

4. Discussion

The results of this study demonstrate that the automated separation of scientific abstracts into user-defined classes might be a technically solvable task. That type of technology might simplify tracking dedicated literature and keeping a researcher up to date, saving a considerable amount of time.

The idea of augmenting a literature search with machine learning was discussed in other papers [2]–[5]. An important benefit of such technologies may be intellectual information retrieval, the prioritization of search results while preparing a systematic review and even future trends prediction [1], [2], [6]. Such systems should be self-tuning and self-updating [7].

A crucial limitation of our work was the relatively small document sizes (article abstracts) used for classifications. Training the models on full texts would probably lead to better results, however, abstracts are incomparably better accessible and are more common for the initial screening in real-world practice [2]. This study is also limited to one user-generated class-balanced dataset, however, a spectrum of machine learning approaches provided additional evidence to support our conclusion. A further improvement of the classification might be related to automatic extraction of in-class intrinsic text features, using the latest state-of-art word embeddings and language models (e.g. BERT), applying ensemble learning and supporting the unbalanced datasets.

5. Conclusions

The classification of scientific publications by their abstracts might be to a certain extent technically solvable and provide a basis for literature tracking in user-defined tasks.

The research was supported by the Russian Foundation for Basic Research grant 19-29-01174.

References

[1] Hana T, Tanaka S, Nejo T, Takahashi S, Kitagawa Y, Koike T, Saito N, Mining-Guided Machine Learning Analyses Revealed the Latest Trends in Neuro-Oncology. Cancers 11 (2019), 178.

[2] Przybyła P, Brockmeier A, Kontonatsios G, Le Pogam MA, McNaught J, von Elm E, Nolan K, Ananiadou S, Prioritising references for systematic reviews with RobotAnalyst: A user study, Res. Synth. Methods 9 (2018), 470–488.

[3] Zaveri A, Cofiel L, Shah J, Pradhan S, Chan E, Dameron O, Pietrobon R, Ang BT, Achieving High Research Reporting Quality Through the Use of Computational Ontologies, Neuroinformatics 8 (2010), 261–271.

[4] Buchlak QD, Esmaili N, Leveque JC, Farrokhi F, Bennett C, Piccardi M, Sethi RK, Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review, Neurosurg. Rev. (2019), 1-19.

[5] Sing DC, Metz LN, Dudli иS, Machine Learning-Based Classification of 38 Years of Spine-Related Literature Into 100 Research Topics, Spine (Phila. Pa. 1976) 42 (2017), 863–870.

[6] Extance A, How AI technology can tame the scientific literature, Nature, 561 (2018), 273–274.

[7] Brockmeier AJ, Mu T, Ananiadou S, Goulermas JY, Self-Tuned Descriptive Document Clustering using a Predictive Network, IEEE Transactions on Knowledge and Data Engineering 30 (2018), 1929-1942.