Augmenting a COVID-19 Research Knowledge Graph With Influential Papers Prediction

Gollam Rabby, Vojtěch Svátek and Petr Berka

[1]Prague University of Economics and Business, Prague, Czech Republic

Abstract

We applied machine learning to predict which of COVID-19-related papers will be highly cited, yielding an extension for the Covid-on-the-Web knowledge graph. Symbolic and deep-learning (BERT) ML performed comparably. LIME-based explanation is also included as part of the produced graph.

Keywords

knowledge graph, COVID-19, research papers, machine learning

1. Introduction

Among the current proliferation of knowledge graphs (KGs), *research-oriented* ones are a particular species. They can be understood as concise, structured representations of various kinds of scholarly knowledge, and have the potential to bridge between overwhelmingly large corpora of scientific texts and the potential recipients of scholarly knowledge who only have limited reading capacity. Numerous projects [1], [2] apply NLP techniques in order to extract key facts from research papers so that they can be exploited independently of their original contexts of publication, without the necessity to read the papers in extenso. The quality of the service provided by the KGs however depends on the quality of papers they represent: knowledge from papers *making impact* in the scientific community should thus be prioritized.

Our present research was motivated by the CIMPLE[1] project, where we envisage a dashboard for fact checkers; one component of it will supply information from KGs relevant for the currently processed claims. Since many claims nowadays relate to COVID-19, we focused our attention to KGs derived from specialized COVID-19 corpora. Of these, *Covid-on-the-Web*[2] is particularly interesting, since it contains argumentative components extracted from papers, which could help sustain the argumentation in the fact check reports. If such a KG were enhanced by paper impact prediction, the fact checkers could more efficiently pick up the argumentation components and possibly also contact the authors for an interview.

We adapted our existing method of CORD-19 paper citation prediction [3], applied it on Covid-on-the-Web, and express its output in RDF so that it could be integrated into this KG.

SEMANTiCS'22, Posters and Demos, September 13–15, 2022, Vienna

*Corresponding author.
rabg00@vse.cz (G. Rabby); svatek@vse.cz (V. Svátek); berka@vse.cz (P. Berka)
0000-0002-1212-0101 (G. Rabby); 0000-0002-2256-2982 (V. Svátek)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://cimple.eu

[2] https://github.com/Wimmics/CovidOnTheWeb

category	Title and Abstract		Title	
	low	high	low	high
Frequency	2541	2538	8063	8058

Table 1
The target variable distribution (discretized citation count adjusted by article age)

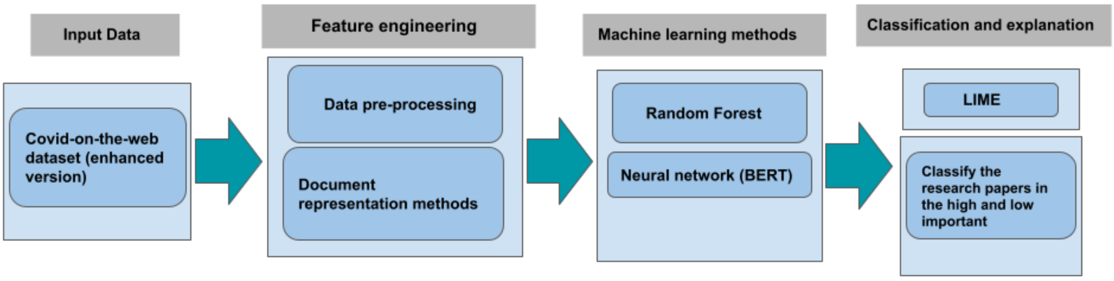
Input Data
Covid-on-the-web dataset (enhanced version)
Feature engineering
Data pre-processing
Document representation methods
Machine learning methods
Random Forest
Neural network (BERT)
Classification and explanation
LIME
Classify the research papers in the high and low important

Figure 1: The data processing pipeline

2. Methods

From our previous experiments we learned that in biomedical scientific document processing, the TF-IDF or bag of words (BOW) representation with random forest or neural network (BERT) learners achieve state-of-the-art results for different combinations of document representation. Also, in most cases, the abstract and title had more impact on classifying a research paper than the bibliometric data had. Therefore we only used the research paper titles and abstracts, for the predictive task.

Input data The *Covid-on-the-Web* KG [4], which describes research papers from biology and medicine relevant to the COVID-19 problem, served as the primary data source for this study. It contains metadata for 16 121 research papers; for 5079 of them it also contains abstracts. We tried to enhance the Covid-on-the-Web KG using from two external datasets, CORD-19[3] and OpenCitations:[4] based on the DOI, we downloaded and cross-checked the data on authors, citation list, citation count, issue, page number, source id, journal name, journal volume, year of publication, title and abstract. The binary target variable for the paper impact classification was derived from the OpenCitation citation counts, using the median value as a threshold. The numbers of papers in impact categories, with regard to the abstract availability, is in Table 1.

The data pre-processing consisted of stopword removal and upper-to-lower-case conversion. (We also experimented with lemmatization and stemming, but their impact on the prediction quality was negligible.)

[3] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
[4] https://opencitations.net/

Data	ML	Repr.		P	R	F1	PCA	AA
Title	NN	BERT	low	0.71	0.80	0.75	0.75	0.74
			high	0.77	0.66	0.73	0.73	
Abstract	NN	BERT	low	0.69	0.64	0.69	0.70	0.72
			high	0.71	0.81	0.76	0.76	
Both	NN	BERT	low	0.77	0.66	0.73	0.73	0.75
			high	0.73	0.81	0.78	0.78	
Title	RF	TF-IDF	low	0.68	0.71	0.70	0.71	0.69
			high	0.68	0.64	0.67	0.67	
Abstract	RF	TF-IDF	low	0.71	0.80	0.75	0.75	0.73
			high	0.76	0.65	0.70	0.70	
Both	RF	TF-IDF	low	0.72	0.81	0.77	0.76	0.75
			high	0.77	0.66	0.72	0.73	
Title	RF	BOW	low	0.71	0.81	0.76	0.76	0.73
			high	0.76	0.65	0.70	0.70	
Abstract	RF	BOW	low	0.71	0.81	0.76	0.76	0.74
			high	0.77	0.65	0.71	0.71	
Both	RF	BOW	low	0.71	0.80	0.76	0.76	0.74
			high	0.76	0.65	0.71	0.71	

Table 2
Prediction quality; Both = both title and abstract; NN = (feed-forward) neural network; RF = random forest; P = precision; R = recall; F1 = F1 score; PCA = per class accuracy; AA = average accuracy.

Document representation The *Term Frequency – Inverse Document Frequency* (TF-IDF) weighting system is the most popular text representation utilized throughout various previous studies. Using the unigrams, bigrams and trigrams from the titles and abstracts we developed the TF-IDF input data table. A binary representation of the input data table (BOW) was also included for comparison, for the same features.

Next, we employed an embedding-based representation approach that is viewed as a cutting-edge within NLP-based language models, *BERT* [5], which was trained on English Wikipedia and BooksCorpus. We used the BERT Tokenizer on the same collection of titles and abstracts.

Machine learning algorithms We used the *random forest* implementation from the scikit-learn library, with the same hyperparameter optimization as by Beranová, et al. [3]. For every input data table (TF-IDF and BOW) the parameters were individually tuned. The focus of the optimization criterion was to improve the accuracy. Next, we used the *simple feed-forward network* from the PyTorch library over the BERT representation.

Explanation algorithm The LIME (Local Interpretable Model-agnostic Explanations) [6] tool demonstrates which feature values and how they affected a certain prediction. This explanation can only be considered approximate because the LIME model is developed by altering the explained instance by varying the feature values and observing the effects on the prediction of each individual feature change. By replacing the described model locally with an interpretable one, the explanation is obtained.

3. Results and Discussion

We used 70% training data and 30% test data by random sampling. The overall accuracy was used to evaluate the results, but we also computed the per-class accuracy. Table 2 shows the Precision, Recall, F1 score and accuracy (per-class and average) of the neural network (BERT) and random forest approach to testing data. As we see, a traditional multi-purpose machine learning algorithms, random forest, performs well like a neural network (BERT). This is not so surprising since also in some other reported cases the difference in performance between BERT, TF-IDF, and BOW was relatively small [7]. Superiority of neural-neural network prediction could possibly be achieved via training domain-driven language models. However, the creation of the TF-IDF and BOW representation is quicker, and the representation enables the use of machine learning techniques that are inherently interpretable while maintaining the interpretability of the generated models.

As regards the RDF output, we store the predicted citation rate category (high or low) together with the citation count from OpenCitations and with the LIME-based interpretation, for every research paper from the Covid-on-the-Web KG. In the GitHub repository[5], the classified data from the covid-on-the-web corpus is available. An example is as follows[6]; the LIME-based explanation (stored just a long string in *lexinfo:explanation*) is displayed as truncated:

```
<https://cimple.vse.cz/covid-on-the-web/10.1016/j.ymeth.2005.05.008>
    a fabio:ResearchPaper , bibo:AcademicArticle , schema:ScholarlyArticle ;
    bibo:doi ”10.1016/j.ymeth.2005.05.008” ; cito:Citation 93 ;
    <https://cimple.vse.cz/covid-on-the-web/expCitationRate> high ;
    lexinfo:explanation ”(’novel’, -0.030867), (’structures’, -0.025789), ...” ;
    schema:url  <https://doi.org/10.1016/j.ymeth.2005.05.008> .
```

4. Conclusions and future work

We have made an initial exploration on augmenting a research-oriented KG with the predicted impact of the underlying papers, obtained via machine learning.

Our next step will be to evaluate this simple approach in the context of a more comprehensive support for users, in particular, the fact checkers, in getting access to scientific literature and its authors. As regards the actual predictive ML technology, the BERT model, having been merely trained on general textual data (English Wikipedia and the BooksCorpus), did not outperform classical ML models in this first try. We however assume that it would work better if trained on domain-specific data such as bio-medical research papers. Also, we also considering to external KGs, such as encyclopaedic ones (DBpedia, Wikidata), into the learning process.

[5] https://github.com/corei5/Enhancement-of-the-Covid-on-the-Web

[6] Prefixes for common vocabularies omitted; they can be retrieved via https://prefix.cc.

5. Acknowledgments

This research is being supported by CIMPLE project (CHIST-ERA-19-XAI-003). The authors also would like to thank Sören Auer and Open Research Knowledge Graph (ORKG) group for providing valuable feedback and some more idea to enhance this with some other research in the ORKG environment.

References

[1] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[2] M. E. Deagen, J. P. McCusker, T. Fateye, S. Stouffer, L. C. Brinson, D. L. McGuinness, L. S. Schadler, Fair and interactive data graphics from a scientific knowledge graph, Scientific Data 9 (2022) 1–11.

[3] L. Beranová, M. P. Joachimiak, T. Kliegr, G. Rabby, V. Sklenák, Why was this cited? explainable machine learning applied to covid-19 research literature, Scientometrics (2022) 1–37.

[4] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, A. Giboin, S. Marro, T. Mayer, et al., Covid-on-the-web: Knowledge graph and services to advance covid-19 research, in: International Semantic Web Conference, Springer, 2020, pp. 294–310.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[6] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[7] M. Mujahid, E. Lee, F. Rustam, P. B. Washington, S. Ullah, A. A. Reshi, I. Ashraf, Sentiment analysis and topic modeling on tweets about online education during covid-19, Applied Sciences 11 (2021) 8438.