Scientometrics (2022) 127:2313–2349
https://doi.org/10.1007/s11192-022-04314-9

Why was this cited? Explainable machine learning applied to COVID-19 research literature

Lucie Beranová[1] · Marcin P. Joachimiak[2] · Tomáš Kliegr[3] · Gollam Rabby[3] · Vilém Sklenák[4,5]

Received: 8 April 2021 / Accepted: 3 February 2022 / Published online: 9 April 2022

Abstract
Multiple studies have investigated bibliometric factors predictive of the citation count a research article will receive. In this article, we go beyond bibliometric data by using a range of machine learning techniques to find patterns predictive of citation count using both article content and available metadata. As the input collection, we use the CORD-19 corpus containing research articles—mostly from biology and medicine—applicable to the COVID-19 crisis. Our study employs a combination of state-of-the-art machine learning techniques for text understanding, including embeddings-based language model BERT, several systems for detection and semantic expansion of entities: ConceptNet, Pubtator and ScispaCy. To interpret the resulting models, we use several explanation algorithms: random forest feature importance, LIME, and Shapley values. We compare the performance and comprehensibility of models obtained by "black-box" machine learning algorithms (neural networks and random forests) with models built with rule learning (CORELS, CBA), which are intrinsically explainable. Multiple rules were discovered, which referred to biomedical entities of potential interest. Of the rules with the highest lift measure, several rules pointed to dipeptidyl peptidase4 (DPP4), a known MERS-CoV receptor and a critical determinant of camel to human transmission of the camel coronavirus (MERS-CoV). Some other interesting patterns related to the type of animal investigated were found. Articles referring to bats and camels tend to draw citations, while articles referring to most other animal species related to coronavirus are lowly cited. Bat coronavirus is the only other virus from a non-human species in the betaB clade along with the SARS-CoV and SARS-CoV-2 viruses. MERS-CoV is in a sister betaC clade, also close to human SARS coronaviruses. Thus both species linked to high citation counts harbor coronaviruses which are more phylogenetically similar to human SARS viruses. On the other hand, feline (FIPV, FCOV) and canine coronaviruses (CCOV) are in the alpha coronavirus clade and more distant from the betaB clade with human SARS viruses. Other results include detection of apparent citation bias favouring authors with western sounding names. Equal performance of TF-IDF weights and binary word incidence matrix was observed, with the latter resulting in better interpretability. The best predictive performance was obtained with a "black-box" method—neural network. The rule-based models led to most insights, especially when coupled with text representation using semantic entity detection methods. Follow-up work should focus on

Extended author information available on the last page of the article

the analysis of citation patterns in the context of phylogenetic trees, as well on patterns referring to DPP4, which is currently considered as a SARS-Cov-2 therapeutic target.

Keywords Bibliometry · CORD-19: COVID-19 open research dataset · Text analysis · SARS-CoV-2 · Interpretability · Citation prediction · Phylogenetic distance · Virus clades

Introduction

With over 50 million research articles written to date (Jinha, 2010), it is often untractable for an individual scientist or group to review all research applicable to the problem at hand. To find documents matching a particular information need, such as checking if a hypothesis has not already been scrutinized or finding prior results to build on, researchers have to rely on specialized search engines. While the search query acts as a first filter, citation count is often the main measure for ordering the documents. Why? As of writing, the only machine-readable information resulting from the many hours researchers spend reading their peers' publications is the number of citations. Despite its limitations, citation count is a useful statistic as several studies have shown it can be used as a proxy for the quality of the article—if the article has many citations, it is more likely to be considered as worth reading.

Predicting the number of citations could be useful, for example, for finding research articles that are "under cited", possibly because they are written by less known authors in specialized or lower impact journals. History shows that such research can sometimes be unnoticed for decades. Gregor Mendel has famously described the laws of inheritance in his paper 'Experiments in Plant Hybridization', which was published in 1866 in Transactions of the Society after being presented at Czechoslovak meetings of Natural History Society in Brünn in 1865. It took 35 years for this seminal work to be rediscovered by biologists in 1900 (Fisher, 1936). More recently, it has been shown that automated analysis of past research literature can be used to identify new materials (Tshitoyan et al., 2019) or potential cancer treatments (Ravanmehr et al., 2021). The issue of timely identification of important research is relevant also in the short-term, such as in the pandemic. When the fast exchange of salient information between scientists is crucial, there is no time to wait for a sufficient number of citations to "naturally" accumulate.

In this article, we attempt to extract some additional insights from the citation count, in addition to the quite general statement about the article's quality. Unlike prior research in bibliometry, which largely focused on analyzing the relationship between article metadata (such as a number of the authors) and citation count, in our research, we focus on semantically interpretable patterns extracted from the article content (including author names), which are predictive of citations.

As the input dataset, we use a subset of articles from the CORD-19 corpus (Wang et al., 2020). Some of the articles in the corpus are recent, while some were published already in the 1970s. We use the number of citations an article has received as a measure of the impact of the research presented in it. In other words, we assume that effective approaches with sound methodology are more likely to get cited.

Goals The hypothesis investigated in our study is that article citations will be associated with combinations of biomedical entities that appear in the text of the articles. We also investigate the applicability of citation biases, such as the preference for authors

with western sounding names. We assume that this type of analysis may help subject matter experts to find new interesting combinations of concepts.

Novelty We are not aware of any systematic research focusing on explainable content-based citation prediction. Many recent works focused on citation analysis (Klavans & Boyack, 2017; Wang, 2018), including multiple works specifically focusing on biomedicine (de Winter, 2015; Kaldas et al., 2020; Rezaee-Zavareh & Karimi-Sari, 2020; Ruano et al., 2018), but none took into account the textual content of the articles. Article text was taken into account only for document clustering (Van Eck & Waltman, 2017), or suggestions of relevant literature (Giosa & Di Caro, 2020). Association rule mining has been recently applied on a semantic representation of the CORD-19 dataset to help uncover frequent patterns in the published research (Cadorel & Tettamanzi, 2020). The limitation of this unsupervised approach is that it does not distinguish articles that made a large research impact from those which did not, and it also relies solely on one modeling algorithm.

A novel element in our study is rule learning from text represented with semantic entities for explainable *classification* of research articles, based on their estimated research impact inferred from the citation count. The bag-of-entities text representation was used in combination with rule learning by us (Kuchař & Kliegr, 2014) as well as by other research teams in combination with neural networks (Yamada & Shindo, 2019), but not in combination with biomedical entity detection or citation analysis.

Methodological contributions The secondary purpose of this research is methodological—to investigate the applicability of a range of existing machine learning and text mining techniques to the problem of content-based citation prediction in general, outside the biomedical domain. The merits of the individual combinations of techniques are assessed based both on their predictive performance and explainability.

Methods

In this research, we adopt the state-of-the-art in natural language processing and use an embeddings-based feature extraction method combined with a neural network classifier, a type of learning algorithm with highly versatile use, including the analysis of textual biomedical data (Mahmud et al., 2020). These high-performance algorithms are coupled with methods for the explanation of "black-box" models.

In parallel, we apply a combination of approaches that are intrinsically explainable ("white-box"). Specifically, this study evaluates a novel combination of entity-based text representation with rule learning. The entity-based text representation detects meaningful biomedical entities (noun phrases). Then, we expand these detected entities with further machine-readable information extracted from a large thesaurus. Rule models, as opposed to neural networks, are intrinsically explainable and can be directly understood by subject domain experts.

In this work, we contrast the intrinsic explanations generated by rule learning algorithms with explanations derived using Explainable Machine Learning techniques from state-of-the-art "black-box" classifiers.

Methodological pipeline In order to obtain generalizable insights, our study employs the following methodology visualized in Fig. 1: (1) Input data collection, (2) Feature engineering, (3) Machine learning (4) Explanation. These steps are described in detail below.

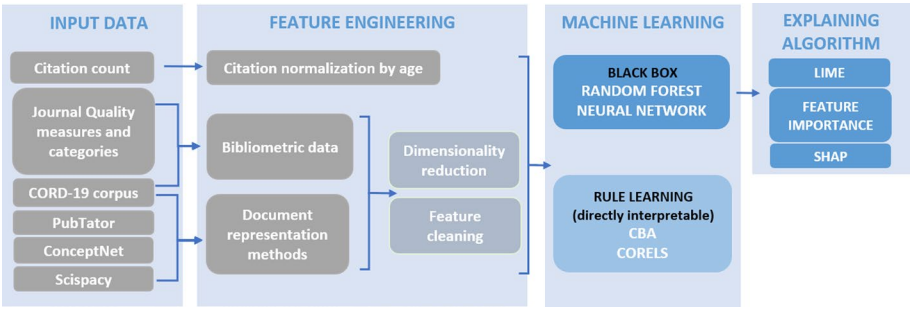

Fig. 1 Overview of methodological pipeline

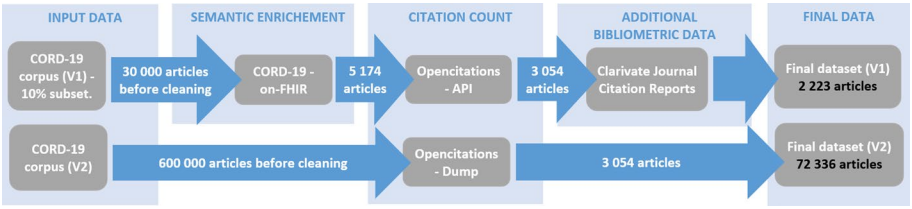

Fig. 2 Process of collecting data and data reduction

Input data

The main data source for this work is the CORD-19 corpus (Wang et al., 2020) containing research articles mostly from biology and medicine applicable to the COVID-19 crisis. The target variable is derived from citation counts, which are not part of the corpus and were obtained by us. We also appended other bibliometric data—journal quality measures and topical categories—for use as additional features.

Input corpus

In our work, we used two versions of CORD-19. As shown in Fig. 2 and as described and justified below, we performed different preprocessing for each of the versions.

Dataset version 1. As the first version of the CORD-19 corpus, we used a release from 2020-10-12 containing approximately 300,000 articles. Since the acquisition of citations and additional entity annotations from external APIs was demanding in terms of time, we limited the number of processed documents by using only the first 10% of articles (30,000) in CORD-19. As a source of additional metadata, we used an additional corpus called CORD-19-on-FHIR[1], which contains semantic annotations mostly generated using Pubtator.[2] By matching "pubmedid" identifiers from CORD-19-on-FHIR, we were able to retrieve metadata for 5174 articles in the used subset of CORD-19. For 2940 of articles with non-empty abstracts, citation counts from the OpenCitations database API were

[1] https://github.com/fhircat/CORD-19-on-FHIR.
[2] https://www.ncbi.nlm.nih.gov/research/pubtator/.

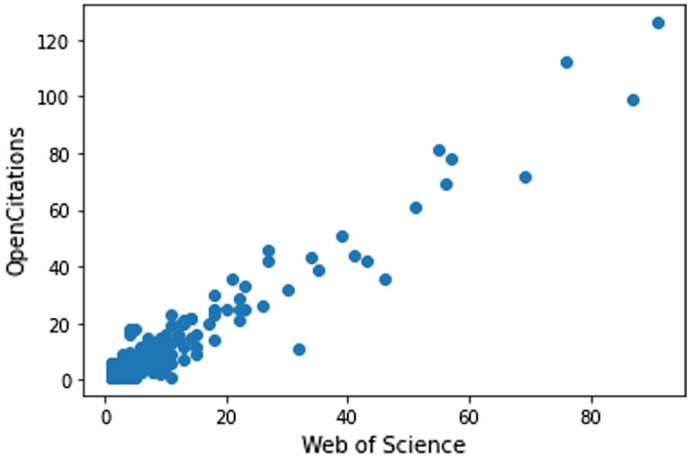

Fig. 3 Correlation between number of citations retrieved from OpenCitations and from Web of Science Expanded API

successfully retrieved. The last reduction of the data resulted from the unavailability of bibliometric data—Journal Quality Measures and Categories for some articles. The final composition of our V1 dataset consisted of 2,223 articles with all necessary information available. All selected articles are in English.

Dataset version 2 To investigate the effect of increasing the dataset size on the accuracy, we checked whether the results based on the smaller sample are sufficiently representative. For this analysis, which was performed as part of a revision of this article, we used a newer release of the CORD-19 corpus (2021-6-22). We also used an up-to-date list of citations, which was retrieved from OpenCitations database dump rather than from the OpenCitations API as in the V1 version. This allowed the processing of a larger number of documents in a timely manner. Note that we have not performed the consequent filtering steps as in V1. In particular, FHIR-to-CORD19 was not updated to match newer CORD-19 versions, and its use would thus excessively reduce the size of the corpus. We also removed articles published in 2021, since for these articles only very limited citation data was available. As a result, the V2 version consisted of 72,336 articles. The distribution of the citations is visualized in Fig. 5.

Bibliometric data—number and quality of citations

We obtained the number of citations based on data from Opencitations.org in order to derive the target variable for the classification model. Before we chose OpenCitations, we compared the citation counts with those retrieved from the proprietary services Microsoft Academic Graph API, Scopus API and Web of Science Expanded API (WoS).

To verify the degree of agreement between data sources, we analyzed citations of 1000 randomly selected articles from CORD-19 (2021-6-22) for which OpenCitations and also WoS citations were available.

With correlation coefficient at 0.97, there is a near-perfect correlation between citation counts retrieved from both data sources (OpenCitations and WoS). The agreement between both data sources for individual articles is visualized in Fig. 3.

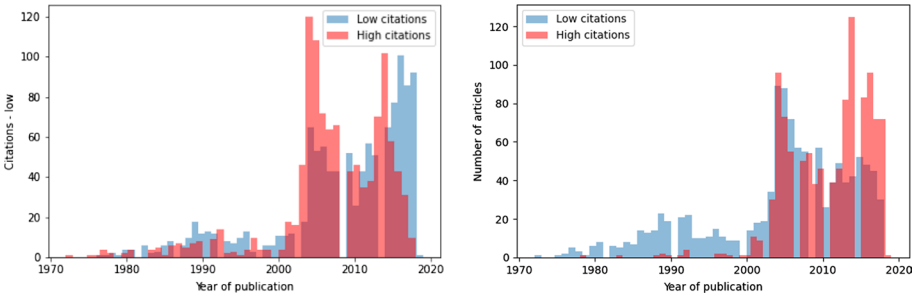

Fig. 4 Distribution of highly vs lowly cited articles before the normalization by age (left) and after normalization (right) for the V1 dataset

In terms of quantity, the average number of citations retrieved from WoS was slightly lower than what we retrieved from OpenCitations (4.33 vs 5.13), which may reflect the curated nature of WoS. We further verified the complementarity of citation sources. To do this, we randomly selected 500 articles from CORD-19 (2021-6-22) for which OpenCitations data were not available. For 60% of these articles (297), citation counts could be retrieved from WoS. The combination of data from multiple citation sources would thus be beneficial, but we left it for future work.

Effect of self-citations and "predatory" citation practices The number of citations of the article should reflect the quality of the scientific research. Because there is no penalty for the excessive number of self-citations, a number of studies suggest caution before accepting self-citations as indicators of scientific impact (Oermann et al., 2020; Soares et al., 2015). Some citation databases try to actively address this problem. In particular, WoS monitors and excludes journals that demonstrate predatory behavior; journals in Journal Citation Reports are subject to additional analysis to detect abnormal citation activity. Journals displaying evidence of excessive self-citation and citation stacking are suppressed from Journal Citation Reports to ensure the integrity of the reports (Web of Science Group, 2022). The high correlation between OpenCitations citation counts (used by us) and WoS citations indicates that the quality of the OpenCitation count are for the purposes of statistical analysis comparable to WoS citations and thus of high quality.

Merging citation counts from multiple sources would probably improve the quality of the results. On the other hand, the inclusion of paid data sources would complicate the replicability of our research. For this reason, we used OpenCitations only.

In the bibliometric literature, a number of schemes for accounting for the age of publications has appeared. In our work, we adopt the proposal from Belikov and Belikov (2015), who suggest dividing the number of citations by the age of the publication in years. The justification provided is that this normalization is adequate since it follows the power-law distribution typical of citations. In our work, the number of citations according to OpenCitations was divided by the number of years, which passed between 2020 and the year when the article was published. The effect of this normalization is depicted in Fig. 4. The plot shows that before normalization, only very few recent articles belonged to the high category, while the normalization corrects this. Notably, only relatively few pre-2000 articles belong to the high category after the normalization. This has a natural explanation since a large number of citations can be attributed to research

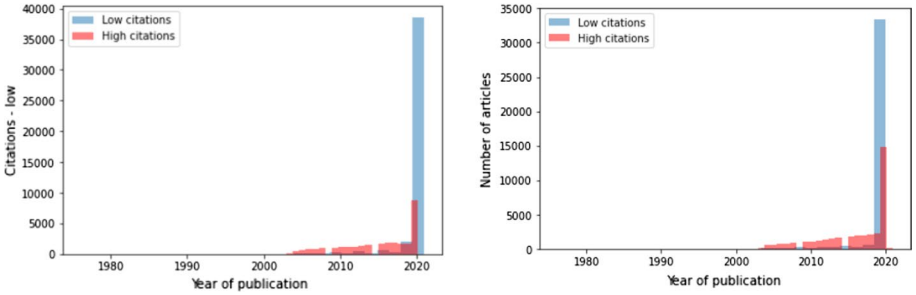

Fig. 5 Distribution of highly vs lowly cited articles before the normalization by age (left) and after normalization (right) for the V2 dataset

Table 1 Distribution of the target variable (discretized citation count adjusted for article age)

Category	V1 (small)		V2 (large)	
	Low	High	Low	High
Citation count	[0;2]	(2;190]	0	(0;2905]
Frequency	1127	1096	36171	36165

related to SARS-CoV-1 and MERS virus outbreaks, which occurred in 2002-2003 and in 2012-2015 respectively (Kumar et al., 2020).

The distribution of the citations in the V2 version is visualized in Fig. 5. As can be seen from comparison with Fig. 4, the use of a newer release of CORD-19 resulted in the addition of a large number of COVID-19 related research articles published in 2020. The last two columns of Table 1 show how the low and high categories were defined to ensure balanced class distribution.

Since some of the state-of-the-art rule learning algorithms support only binary target variables, we transformed the citation count into a categorical predicted variable. We thus cast the problem as a classification task. We also considered directly formulating the problem as a regression task predicting a specific number of citations and then applying thresholding. However, as shown in our results section, this approach was less successful.

We thus performed equifrequent binning into two categories (bins). The use of the equifrequent algorithm resulted in nearly perfectly balanced datasets, with both target classes having almost the same number of articles (Table 1).

In our initial experiments, we performed binning into three manually designed categories of three logarithm 10 bins (< 10 citations, [10;100), >100 citations). As follows from earlier bibliometric studies, there are many more articles with a low number of citations than there are highly cited articles (Vieira & Gomes, 2010). As a consequence, the resulting classes were unevenly populated, requiring an application of oversampling or undersampling approaches such as SMOTE (Chawla et al., 2002), which would make both analysis and explanation more convoluted. Therefore, after this initial study, we decided to use only two target categories and leave a more complex binning for future work. Since all compared algorithms use the same input, this should not substantially affect the fairness of the comparison. Moreover, the two created categories have clear semantics. The *low* category with up to two citations (per year—due to normalization) corresponds to articles that

were subject of very modest follow-up interest, while the *high* category is for articles that were cited and built upon by other researchers.

Other bibliometric data as features

Journal quality measures and categories Prior bibliometric research focused on more general article features has revealed that citation count is positively correlated with the number of authors, a longer length of the article, a higher number of references, and particularly journal impact factor (Vieira & Gomes, 2010). The aforementioned research found these correlations to hold also specifically for biology and biochemistry, which are the major domains in the CORD-19 corpus.

Since the bibliometric study of Vieira and Gomes (2010) has been published, the field of bibliometry has adopted Article Influence Score (AIS) as a more robust replacement for the impact factor (Roldan-Valadez et al., 2018). We used Clarivate Journal Citation Reports (JCR)[3] as the source of both the impact factor and AIS.

We also used this database to include the information on the primary (first) topical category to which the journal belongs, such as "Virology". We also retrieved a less granular categorization via UNESCO's Fields Of Research And Development (FORD) taxonomy, which maps Virology to the "Biological Sciences" category.

It should be noted that while the mapping between CORD-19-on-FHIR documents and JCR was straightforward in most cases, for more than one thousand articles, the listed journal names or abbreviations did not match the ISO-standardized journal name or abbreviation in JCR. In these cases, we performed automated matching based on text similarity. A relatively small number (711) of CORD-19 articles, which could not be mapped even with this method—mostly due to a completely missing journal title—was removed from the V1 dataset.

Feature engineering

The first set of features was derived from the bibliometric indicators and placed into the "Bibliometric data" matrix. This included the following information:

- number of authors,
- license (such as CC0, CC-BY, CC-BY-NC, bioarxiv license),
- impact factor and article influence score for a year when the article was published, when unavailable, the first known impact factor for the journal since the article was published,
- latest impact factor and article influence score (2019),
- first JCR category and its FORD mapping,
- tokenized journal name.

From the CORD-19-on-FHIR dataset, we used the abstracts of the articles. We decided not to use the full-texts, since this would limit the applicability of our method as well as reproducibility of our results as unlike full texts, abstracts are almost always available. The abstracts were processed in three alternative ways: using the standard TF-IDF approach,

[3] https://jcr.clarivate.com.

Table 2 Entity recognition systems used

Training corpus	Entity types
CRAFT	GGP, SO, TAXON, CHEBI, GO, CL
JNLPBA	DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN
BC5CDR	DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN
BIONLP13CG	AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE

List adapted from https://allenai.github.io/scispacy/

binary word incidence matrix, and embeddings. Additional feature matrices were obtained by applying entity detection and semantic expansion methods. Finally, dimensionality reduction methods have been applied. The details are covered in the following "Document representation methods" section.

Document representation methods

TF-IDF matrix Possibly the most common text representation used in prior work in bibliometry is the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme, cf. e.g., Glenisson et al. (2005). The TF-IDF matrix was created for unigrams, bigrams, and trigrams extracted from article abstracts.

Binary bag-of-words incidence matrix (BOW) The reason for including the binary version of the BOW matrix is that the rule learning algorithms that we utilized are able to learn only from binary features. The application of discretization on TF-IDF scores and subsequent binarization (e.g., via one hot encoding or dummification) would be possible, but it would substantially increase the already high dimensionality and sparseness of the document-term matrix. Similarly, as the TF-IDF matrix, the BOW matrix was created for unigrams, bigrams, and trigrams extracted from article abstracts.

Embeddings with BERT As a representative of the state-of-the-art embeddings-based approach, we used the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). We applied the BERT Tokenizer on the same set of abstracts in the CORD19-on-FHIR corpus as the previous two document representation techniques. We used a pretrained model[4] with twelve hidden layers, the hidden size of 768, and twelve attention heads. The weights were the same as released by the original model authors (Devlin et al., 2019) and the pretraining was performed on English Wikipedia and BooksCorpus (Zhu et al., 2015).

While new language models, such as BERT used in our work, provide excellent predictive performance on some tasks (Devlin et al., 2019), the difference in performance

[4] https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/2.

between BERT and TF-IDF is sometimes reported as relatively small (Piskorski et al., 2020). On the other hand, generating the TF-IDF representation is faster and the resulting vectors allow for the application of intrinsically interpretable machine learning algorithms, retaining the interpretability of the resulting models. *Entity extraction* For entity extraction, we used a transition-based system based on the chunking model (Lample et al., 2016) as implemented in ScispaCy (Neumann et al., 2019). We used four pretrained ScispaCy models depicted in Table 2.

All models were executed independently, and the entities detected were merged into one feature vector. The feature vector contained only the detected entities. We did not use entity types, as these are not supported by all ScispaCy models, and when they are supported, they appear too general for our purpose as can be seen in Table 2. The most specific detected types are CELL or GENE, which is too coarse-grained. To construct the feature vector, we used the lemma returned by ScispaCy, which also made the semantic expansion covered in the next paragraph more successful.

Entity Expansion with ConceptNet Once entity extraction has been performed, we performed entity expansion by adding related concepts. These were retrieved using the ConceptNet semantic network (Speer et al., 2017). The semantic relations covered by ConceptNet include synonyms in other languages (for entity polypeptide, this is, e.g., peptidi in Finnish), related terms (polypeptidase, …), subtypes (adrenocorticotrophin,…) , supertypes (peptide, polymer, …), derived terms (copolypeptide, …) and context (organic chemistry, protein).[5] As can be seen in this example, while ConceptNet is a general knowledge network, it also covers concepts related to specialized knowledge in CORD-19.

Input to ConceptNet expansion was a list of entities E_d detected in a given input document d as described earlier. As part of entity expansion, for each document d, we identified a set of related ConceptNet entities C_d so that for each $c \in C_d$, ConceptNet 5 contains an edge $e \rightarrow c$, or $c \rightarrow e$, where c is a ConceptNet entity and e is a ConceptNet entity detected in document d, $e \in E_d$. Denoting the set of all ConceptNet expansion entities as

$$C = \bigcup_{i=1}^{N} C_i, \tag{1}$$

and the number of documents as N, we created a binary ConceptNet matrix M of dimension $N \times |C|$, where $M_{jk} = 1$ if and only if at least for one entity detected in document j there is an edge between this entity and entity k in ConceptNet.

PubTator annotations The CORD-19-on-FHIR dataset comes with annotations pregenerated using Pubtator (Wei et al., 2013), which is a system for generating automatic annotations of biomedical concepts. Each article was represented using text node annotations present in the source dataset. We considered only text nodes that were shorter than 40 characters and did not contain any non-ASCII character. Example generated extracted annotations for one article included *infection*, *viral infection*, *human*, *CHME-5*, *astrocytoma*, *murine*, and *oligodendrocytic*. Annotations longer than 20 characters tended to contain multiword fragments of text, which did not correspond to a single entity. To generate the binary PubTator feature matrix, documents were represented in the same way as described in the previous section for ConceptNet.

[5] https://conceptnet.io/c/en/polypeptide.

Table 3 Overview of input datasets

Dataset (Matrix)	Features	Reduction method	Columns	Original columns
AuthorsNames	Binary	min_df = 32	162	504,237
BibliometricFeatures	Binary	min_df = 32	539	504,614
Bow	Binary	min_df = 32	1495	547,769
Bow_BibliometricFeatures	Binary	min_df = 32	2034	1,052,383
TF-IDF	float	min_df = 32	1495	547,769
TF-IDF_BibliometricFeatures	mixed	min_df = 32	2034	1,052,383
PubTator	Binary	FI = max 1500	1500	3322
PubTator_Conceptnet	Binary	FI = max 1500	1500	2087
ScispaCy	Binary	FI = max 1500	1500	23,685
ScispaCy_Conceptnet	Binary	FI = max 1500	1500	9483
Bow_PubTator	Binary	min_df = 32, FI = max 1500	2995	551,091
Bow_PubTator_Conceptnet	Binary	min_df = 32, FI = max 1500	2995	549,856
Bow_ScispaCy	Binary	min_df = 32, FI = max 1500	2995	571,454
Bow_ScispaCy_Conceptnet	Binary	min_df = 32, FI = max 1500	2995	557,252
Bow_Pubtator_Conceptnet _BibliometricFeatures	Binary	min_df = 32, FI = max 1500	3534	1,054,470

Table 4 Evaluation of BOW matrix for different value of the minimum document frequency (min_df) parameter with the RandomForest classifier

min_df	Features	Fit time	Accuracy
1	426,272	87.03	0.72
4	13,397	4.25	0.72
8	5848	1.86	0.73
12	3841	1.23	0.73
16	2940	0.97	0.73
20	2359	0.83	0.72
24	1961	0.72	0.73
28	1708	0.66	0.72
32	1495	0.60	0.73
36	1350	0.57	0.71
40	1178	0.54	0.72
44	1079	0.51	0.72
48	985	0.49	0.72
52	894	0.47	0.73
90	456	0.36	0.71
120	320	0.32	0.71
150	232	0.30	0.70
200	153	0.28	0.70
250	101	0.26	0.70

Dimensionality reduction of input datasets

An overview of the datasets representing the individual sets of features and their combinations is provided in Table 3. This table also includes the results for various thresholds for feature selection, which was applied to reduce the size of the dataset to address scalability issues encountered with the rule learning algorithms. In order for the rule models to be learnt in a reasonable time (less than 1 hour), the input matrices have been reduced. For the (binary) BOW matrix and the TF-IDF matrix we excluded terms with low document frequency, similarly as, e.g., in (Piskorski et al., 2020). To set the threshold value, we investigated the relationship between the dimension of the matrix and the accuracy of the random forest model trained on it. Table 4 shows that in the case of the BOW matrix, the highest accuracy of 73% is most stably attained for a vector length of about 3000. Other matrices, like ScispaCy, ScispaCy Conceptnet, PubTator, and PubTator Concepnet, have also been reduced. We used a maximum of 1500 most important variables according to the MDI feature importance scores (cf. "Explanation algorithms" section). Other matrices are created by merging the existing ones.

Feature cleaning for improving interpretability

For rule learning, where the individual features are shown to the user as part of the rules, we also experimented with additional preprocessing of the Bow_Pubtator_Conceptnet matrix, which merges the text-based features from three sources and is used as a basis for interpreting rule learning results. For rule learning purposes, we generated a variant version of this matrix with additional feature cleaning as described below.

Unification of features For the original Bow_Pubtator_Conceptnet matrix, in some cases clashing features were generated, i.e., the same feature appeared in each of the composite matrices, but the set of documents in which this feature was detected differed. For example, entity 'cats' was detected both as a word in the text and as an entity in the Pubtator annotations. By default, we kept such entities as separate features differentiated by a suffix. However, in the matrix version with extra cleaning, these features were collapsed into one feature, which was set to 1 if any of the composite features was 1.

Stopword removal Removal of stopwords did not affect the classifier accuracy but improved interpretability, therefore this was used in the version with extra cleaning.

Lemmatization and stemming We attempted replacing features with their lemmas, as well as stems. Both techniques decreased classifier accuracy, therefore this preprocessing was not used in either version.

Machine learning algorithms

As modeling algorithms, we tried to use a representative selection of current approaches—random forests and neural networks. According to multiple benchmarks, these algorithms provide the best performance on a wide variety of different tasks (Fernández-Delgado et al., 2014; Wainberg et al., 2016). As representatives of the rule learning approach, we used Certifiably Optimal Rule Lists (CORELS) introduced by Angelino et al. (2017) and Classification Based on Associations (CBA) introduced by Liu et al. (1998). Both approaches are

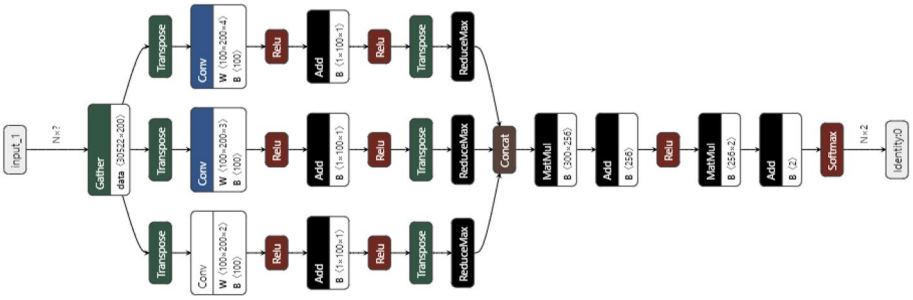

Fig. 6 Architecture of the used Convolutional Neural Network, generated by Netron (https://netron.app/)

Table 5 Hyperparameter combinations evaluated for the neural network model

Parameters	Run 1	Run 2	Run 3	Run 4
EMB_DIM	200	1400	500	1300
CNN_FILTERS	100	130	200	50
DNN_UNITS	256	256	256	256
OUTPUT_CLASSES	3	3	3	3
DROPOUT_RATE	0.2	0.2	0.2	0.2
NB_EPOCHS	5	5	5	5
Accuracy	0.56	0.70	0.59	0.68

representatives of associative classification, which has been used by Iqbal et al. (2020) for a similar task — extracting combinations of writing style features for authorship attribution.

In the following, we describe the setting of individual algorithms in detail.

Random Forest We used the Random Forest implementation from scikit-learn.[6] Hyperparameter optimization was performed by grid search as part of internal cross-validation during model learning. The model parameters were optimized separately for each input matrix. The optimization criterion was accuracy. The parameter grid was as follows:

- 'max_depth': {10,150,500,1000},
- 'max_features': {30,500,3000},
- 'min_samples_leaf': {1,10,100},
- 'min_samples_split': {2,10,100},
- 'n_estimators': {10, 100}

*Neural Networks (over BERT)*Lee and Dernoncourt (2016) point out the effectiveness of convolutional neural network (CNNs) compared to other network architectures (LSTMs and Recurrent Neural Networks) for short text classification. Our model consists of three CNN layers as visualized in Fig. 6. The model was learned for four combinations of hyperparameters as depicted in Table 5.

CORELS We used the CORELS implementation from the method's author[7] with the default parameters. The maximum number of nodes was set to 100,000, and regularization

[6] https://scikit-learn.org/.

[7] https://github.com/corels/pycorels.

strength was set to 0.01. This value corresponds to adding a penalty equivalent to misclassifying 1% of instances when adding the additional rule to the list of the generated rules. We also experimented with other regularization settings (including disabling regularization by setting the corresponding parameter to 0), but the effect on the resulting model was small. The minimum support bounds optimization and lookahead bound optimization were enabled.

CBA We used our implementation of CBA, which is available in the `arc` R package (Hahsler et al., 2019). We used the recommended hyperparameter values as suggested by the author of the method in (Liu et al., 1998): minimum support of 1%, minimum confidence 50%. In an evaluation of the effect of tuning CBA hyperparameters reported by Kliegr and Kuchař (2019), it was found that different sets of these thresholds do not noticeably improve predictive performance therefore we did not perform any tuning. We additionally applied a limit on the rule length of four items.

Explanation algorithms

The algorithms used for modeling include rule learning, which generates a directly interpretable representation, as well as random forests and neural networks, whose interpretation requires the application of additional explanation algorithms. In the following, we provide a brief overview of these explanation approaches.

Rule models Both CORELS and CBA algorithms involved in our study generate *rule lists*. A rule list is an ordered collection of rules, where each rule is associated with a distinct priority value. A rule has a form of the *antecedent* → *consequent*. The rule consists of a set of conditions (antecedent). The consequent contains the predicted value of the target class. To classify (predict a class) for a particular instance, the evaluation algorithm processes the rules in the rule list in the order of priority, highest to lowest. Once it finds a rule with all conditions in the antecedent matching the current instance, the consequent of this rule is used as a prediction for the instance. Rules with lower priority are not processed.

The principal difference between CORELS and CBA is the type of models they produce. The CORELS algorithm tends to provide very condensed models, often containing only one if-else rule. In this respect, the application of the CORELS algorithm on real data for parole and bail decisions received significant attention. In a paper titled *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead* (Rudin, 2019), some of the authors of CORELS have argued that a rule model composed of several rules generated by CORELS performs comparably in terms of accuracy to the black-box COMPAS model widely used for actual bail decisions in the U.S. Angelino et al. (2017) have also shown that CORELS is competitive against multiple other machine learning models, including C4.5 and CART decision trees.

CORELS outputs very short models, which is not always desirable. As also demonstrated in our experiments, CORELS models can have lower predictive performance. To complement CORELS, we used CBA, which is one of the most commonly used rule learning algorithms based on association rule learning. Unlike CORELS, CBA results in models that contain a higher number of rules. The advantage of this approach is that it provides better insight into the data since the individual rules correspond to local patterns and could

thus be used as a tool for descriptive data mining or explanation. The disadvantage is that CBA-generated models may contain too many rules for the user to be able to manually review. To present these rules in a concise way, we applied grouped matrix clustering, which is a rule clustering technique recently proposed by Hahsler and Karpienko (2017). To communicate the shared elements between the rules, we used the graph-based rule visualization, also adapted from Hahsler and Karpienko (2017).

Random forest and neural networks Both these algorithms belong to the group of "black-box" approaches, which are characterized by the opacity of the internal working of the generated models.

Models generated from random forests cannot be directly interpreted due to the number of trees, their complexity and also the fact that multiple trees can take part in the decision. However, the random forest learning algorithm was designed so that estimates of feature importance scores are readily provided (Breiman, 2001). In our work, we adopt the original method for computing the feature importance scores of random forests, which is based on Mean Decrease of Impurity (MDI). For this method, it has been shown that the MDI importance of a relevant feature is invariant with respect to the removal or addition of irrelevant features and that the importance of a feature is zero if and only if the feature is irrelevant (Louppe et al., 2013).

For neural networks, a number of feature importance methods have been proposed, but as has been recently shown, many of these methods do not provide stable results (Ghorbani et al., 2019). In addition, there is a number of model agnostic methods by which feature importance for models like random forests can be computed. In our work, we adopt Shapley values (Lundberg et al., 2020) and LIME (Ribeiro et al., 2016). Unlike the MDI method for Random Forests, which generates global feature importance scores, these algorithms provide local feature importance values for a particular test instance.

SHapley Additive exPlanations (SHAP) value emerges from the Shapley concept from game theory (Rodríguez-Pérez & Bajorath, 2020). The SHAP values allow global interpretation. Each observation gets its own set of SHAP values so it is possible to also interpret it locally.

LIME (Local Interpretable Model-agnostic Explanations) shows which feature values contributed to a particular prediction and how. This explanation is only approximate since the LIME model is learnt by modification of the explained instance by perturbing the feature values and collecting the resulting impact of each individual feature change on the prediction. The explanation is obtained by locally approximating the explained model with an interpretable one.

Results

We provide two perspectives—predictive performance and model interpretability. When interpreting the models, we noticed a pattern indicating that low citations are associated with Asian names and high citations more commonly with western-sounding names. The last part of this section is devoted to detailed results related to this phenomenon.

Predictive performance

To evaluate predictive performance, we split the data into 70% training and 30% testing. The results were evaluated in terms of accuracy, computed as the number of correct

Table 6 Predictive performance of random forests and neural networks for V1 dataset of 2223 articles

Matrix	Binary	Regression	
	Accuracy	MSE	Accuracy
Random Forest			
AuthorsNames	0.68	25,25	0.55
BibliometricFeatures	0.68	21.26	0.61
Bow	0.70	22.79	0.61
Bow_BibliometricFeatures	0.70	21.98	0.59
Bow_PubTator	0.72	22.50	0.61
Bow_PubTator_Conceptnet	0.71	22.66	0.62
Bow_PubTator_Conceptnet_Bibliometric_Features	0.72	17.61	0.67
Bow_ScispaCy	0.69	22.61	0.59
Bow_ScispaCy_Conceptnet	0.70	22.36	0.64
PubTator	0.68	27.90	0.60
PubTator_Conceptnet	0.67	28.60	0.58
ScispacC	0.60	32.12	0.53
ScispaCy_Conceptnet	0.60	31.80	0.53
TF-IDF	0.70	25.17	0.54
TF-IDF_Bibliometric_Features	0.71	21.30	0.61
BERT embeddings	0.67	32.92	0.54
Neural Network			
BERT embeddings[a]	0.83	22.04	0.80

[a]BERT results were updated for the final version of the article using BERT TF HUB Model (bert_en_uncased_L-12_H-768_A-12/2) instead of V1 of the same model. The previous accuracy for BERT (Classification) was 0.71 and accuracy for BERT (regression) was 0.56

classifications divided by the number of all predictions. We have also evaluated the model's F1 score, but we do not report this since it was in most cases almost identical (within 2%) to the model accuracy. We attribute this to the fact that the dataset was balanced.

The main results for the V1 version of the dataset are presented in Table 6 for "black-box" model (random forest and neural network) and in Table 7 for "white-box" models generated with rule learning (CORELS and CBA).

We refer to results obtained with Random Forest as the main baseline. The first interesting result following from Table 6 is that the binary BOW matrix performs equally well as the matrix with TF-IDF weights, despite the fact that it contains more information. All attempts at representing the abstracts only with extracted entities (ScispaCy, PubTator) resulted in a lower predictive accuracy than this baseline. Further expansion with ConceptNet had no effect either. As expected, the matrix combining all available features for random forest had the best predictive performance with 72% accuracy.

The combination of BERT document representation with Neural Network training performs better than RandomForest trained on the TF-IDF vectors.

Results for rule learning algorithms are summarized in Table 7. The state-of-the-art CORELS classifier did not match the predictive performance of the CBA algorithm. Overall, there is about 4% difference between the best Random Forest model trained over BOW-based representation and the best rule-based model. This gap can be attributed to

Table 7 Predictive performance and model size of rule learning (CBA and CORELS) for V1 dataset of of 2223 articles

Matrix	CORELS			CBA		
	Accuracy	avgRuleLen	ruleCount	Accuracy	avgRuleLen	ruleCount
AuthorsNames	0.51	1.0	1	0.67	1.3	99
BibliometricFeatures	0.66	1.5	2	0.69	2.1	192
Bow	0.64	1.5	2	0.66	2.2	350
Bow_BibliometricFeatures	0.65	1.5	2	0.68	2.1	417
Bow_PubTator	0.66	1.5	2	0.67	2.1	349
Bow_PubTator_Conceptnet	0.66	1.5	2	0.67	2.1	424 (465)
Bow_Pubtator_Conceptnet _BibliometricFeatures	0.61	1.5	2	0.68	2.0	349
Bow_ScispaCy	0.61	1.5	2	0.68	1.5	162
Bow_ScispaCy_Conceptnet	0.65	1.0	2	0.67	1.1	121
PubTator	0.62	1.8	5	0.66	1.7	64
PubTator_Conceptnet	0.64	1.5	2	0.64	1.9	73
ScispaCy	0.60	1.5	2	0.57	0.5	2
ScispaCy_Conceptnet	0.60	1.5	2	0.57	0.7	3

For Bow_PubTator_Conceptnet, the number in parenthesis is for the version with extra feature cleaning described in "Feature cleaning for improving interpretability" section, the remaining results were the same as for the base version

the interpretability-accuracy trade-off and is several percentage points lower than the average difference between the accuracy of the best Random Forest model and the best rule-based model as reported in (Fernández-Delgado et al., 2014). Somewhat unexpectedly, rule learning (CBA) slightly outperformed Random Forest on the Bibliometric features dataset. However, the gap between the best rule-based model and a neural network trained over the BERT representation is substantially bigger.

Effect of training a regression model instead of a binary classification model

To evaluate the benefit of the early transformation of the problem into a binary classification task, we trained a regression model using the V1 datasets, with the target variable being the citation count. We evaluated the regression model using Mean Square Error (MSE). Then, we applied the same threshold as in our main analysis but on the *predicted* citation counts. The results are shown in the last two columns of Table 6. This experiment shows that better accuracy is obtained when the problem is reformatted to a binary classification problem as opposed to when the problem is dealt with using regression with subsequent thresholding.

Effect of the training data size

Here we investigate the effect of increasing the dataset size on the accuracy of the models. For this analysis, we used the V2 version of the dataset, which was with nearly 72,238 articles more than 30× larger than V1.

Table 8 Topmost important features (MDI method) by matrix

BibliometricFeatures		PubTator_Concepnet		Bow_PubT_Conc_BiblFeatures	
Feature	Imp.	Feature	Imp.	Feature	Imp.
FORD_0_impactQ_Q2	0.076	Mers	0.050	FORD_0_impactQ_Q2	0.026
FORD_0_aisQ_Q1_D1	0.047	Humans	0.023	east	0.021
WoScategory_0_impactQ_Q3	0.031	Human	0.021	middle east	0.019
FORD_0_impactQ_Q1_D2	0.031	Dromedary	0.019	east respiratory	0.017
license_elscovid	0.029	Camels	0.016	WoScategory_0_aisQ_Q1_D2	0.016
FORD_0_aisQ_Q2	0.029	Cov	0.015	FORD_0_aisQ_Q1_D1	0.015
WoScategory_0_aisQ_Q3	0.028	Cow	0.013	FORD_0_impactQ_Q1_D2	0.014
FORD_0_impactQ_Q1_D1	0.028	Body	0.010	east respiratory syndrome	0.013
WoScategory_0_aisQ_Q1_D2	0.025	Infection	0.009	respiratory syndrome	0.012
FORD_0_aisQ_Q1_D2	0.023	Rats	0.009	syndrome	0.010
license_unk	0.022	Fever	0.008	license_unk	0.010
WoScategory_0_aisQ_Q1_D1	0.020	Bovine	0.008	FORD_0_aisQ_Q1_D2	0.009
FORD_0_aisQ_Q3	0.013	Canine	0.008	WoScategory_0_aisQ_Q3	0.009
peter	0.011	Failure	0.008	middle_east_raspiratory	0.009
journal_Journal_of_Virology	0.010	C	0.007	middle	0.008
journal_Arch_Virol	0.008	Pneumonia	0.007	WoScategory_0_aisQ_Q1_D1	0.008
WoScategory_0_impactQ_Q2	0.008	Respiratory	0.007	FORD_0_impactQ_Q1_D1	0.007
WoScategory_0_aisQ_Q2	0.008	Transgenic	0.006	FORD_0_aisQ_Q2	0.007
WoScategory_0_impactQ_Q2	0.007	Dog	0.006	journal_Journal_of_Virology	0.007
paul	0.007	People	0.006	license_elscovid	0.005

(WoScategory|FORD)_0 indicates the value is for the journal's primary FORD (Web of Science) category. (AIS|Impact)Q_Q{q} indicates that the journal in which the publication appeared is in the q-th quartile by AIS (impact factor). If the journal is in the first two deciles of Q1, then D{d} indicates the decile

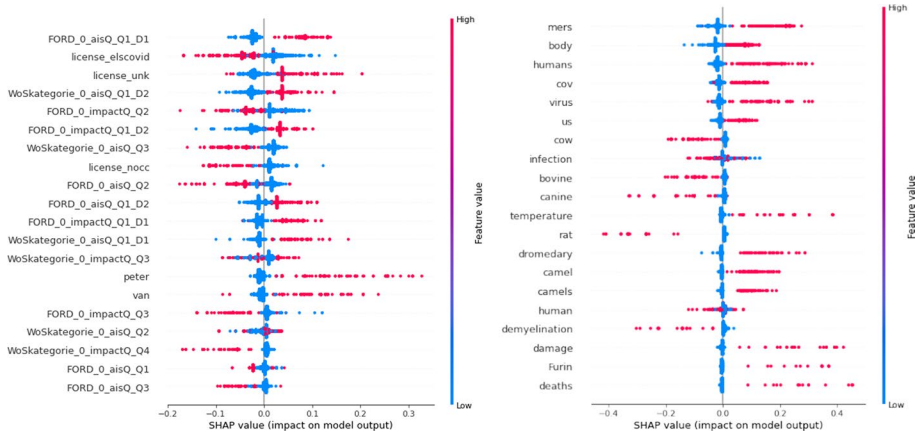

Fig. 7 Shapley plot for bibliometric features (left) and article abstracts (right). Features are sorted by mean SHAP value. Example explanation: articles annotated with *license_nocc* "no Creative Commons license" in CORD-on-FHIR-19 have value 1 (denoted by red dots), and articles with other license value 0 (blue dots). Concentration of red dots left of the vertical line (SHAP value < 0) indicates that article license "nocc" has a negative effect on the number of citations. Note that some features like camel and camels could have been aggregated by stemming. This was not performed for the Random Forest model, since it had negative effect on predictive performance (e.g. human and humans are often used in different contexts). (Color figure online)

Table 9 Predictive performance of Random Forest and Neural Network for both versions of the input dataset

Model	Matrix	Accuracy	
		V1 (small)	V2 (large)
Random Forest	Bow	0.70	0.70
Random Forest	TF-IDF	0.70	0.70
Neural Network	BERT	0.83	0.66

Results for V1 are taken for reference from Table 6 (2223 articles), V2 dataset contains 72,336 articles

Fig. 8 LIME plot for authors

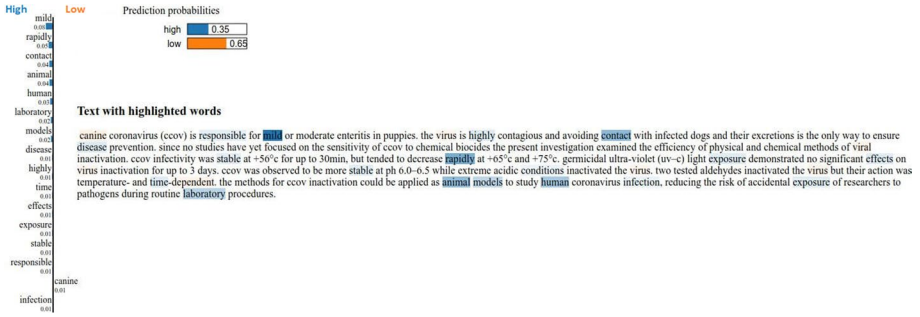

Fig. 9 LIME plot for abstract

The results presented in Table 9 indicate that the availability of the additional training data had no effect on the accuracy of Random forest and had a small negative effect for the Neural Network.

Model interpretation

Random forest models can be interpreted through feature importance values. Top features by importance computed by the MDI method for three representative matrices are presented in Table 8. The importance of individual bibliometric features and abstracts is captured using a SHAP plot in Fig. 7. Note that, unlike the MDI feature importance values, the SHAP plot also captures the direction of the effect. LIME plot explaining the prediction of a random forest model for a representative document based on its author information is in Fig. 8 and based on the text of the abstract in Fig. 9.

Summary statistics for rule learning models are present in Table 7. Remarkably, for datasets containing entities extracted from PubTator, which were further semantically

Table 10 Example rules generated by the CBA algorithm grouped by input dataset (matrix)

Matrix	LHS	RHS	Supp.	Conf.	Cov.	Lift
Bow_ScispaCy	{oc43 strain,bcv}	{low}	22	1.00	0.01	1.89
Bow_ScispaCy	{merscov infection,virus}	{high}	22	1.00	0.01	2.13
Bow_ScispaCy	{killing,merscov infection}	{high}	44	0.97	0.02	2.07
Bow_ScispaCy	{neutralizing antibody, merscov infection}	{high}	44	0.96	0.02	2.04
Bow_ScispaCy_Conc	{bradycardie,bcv}	{low}	22	1.00	0.01	1.89
Bow_ScispaCy_Conc	{results indicated}	{low}	23	0.96	0.01	1.81
Bow_ScispaCy_Conc	{canine,dogs}	{low}	21	0.96	0.01	1.81
Bow_ScispaCy_Conc	{merscov infection}	{high}	40	0.94	0.03	2.01
Bow_PubTator	{reversed,bcv}	{low}	22	1.00	0.01	1.89
Bow_PubTator	{merscov infection}	{high}	40	0.94	0.03	2.01
Bow_PubT_Concepnet	{canine,virus}	{low}	23	1.00	0.01	1.89
Bow_PubT_Concepnet	{bats,coronavirus,transmission}	{high}	17	1.00	0.11	2.01
Bow_PubT_Concepnet	{merscov,mice}	{high}	22	1.00	0.01	2.13
Bow_PubT_Concepnet	{ifn,innate,respiratory}	{high}	22	1.00	0.01	2.13
Bow_PubT_Concepnet	{mice,protection,vaccine}	{high}	22	1.00	0.01	2.13
Bow_PubT_Concepnet	{dpp4,respiratory}	{high}	22	1.0	0.014	2.02
Bow_PubT_Concepnet	{virus,infectious}	{high}	34	0.72	0.02	1.53
BibliometricFeatures	{FORD_0_aisQ_Q1_D1 ,christian}	{high}	20	1.00	0.01	2.02
BibliometricFeatures	{FORD_0_impactQ_Q1 _D1,van}	{high}	22	1.00	0.01	2.02
BibliometricFeatures	{WoSkateg._0_obor_ VIROLOGY_SCIE, FORD_0_aisQ_Q1_D2}	{high}	378	0.7	0.24	1.41
BibliometricFeatures	{FORD_0_ford_10600, FORD_0_aisQ_ Q1_D2, FORD_0_impactQ_Q1_D2}	{high}	375	0.7	0.25	1.4
Bow_BibFeatures	{FORD_0_aisQ_Q1_D1, merscov}	{high}	67	1.00	0.03	2.02
Bow_BibFeatures	{antibodies,middle east}	{high}	44	1.00	0.03	2.02
Bow_BibFeatures	{WoScategory_0_ aisQ_Q1_D1,merscov}	{high}	67	1.00	0.03	2.02
Bow_BibFeatures	{middle east,spike protein}	{high}	36	1.00	0.02	2.02
Bow_BibFeatures	{dromedary, east respiratory syndrome}	{high}	35	1.00	0.02	2.02
Bow_BibFeatures	{WoScategory_0_obor_ VIROLOGY_SCIE, homology,sequence}	{low}	22	0.89	0.01	1.76
Bow_BibFeatures	{FORD_0_impactQ_Q2, coronavirus, substitutions}	{low}	22	0.89	0.01	1.76
Bow_PubT_Conc_BibF	{merscov, FORD_0_aisQ_Q1_D1}	{high}	67	1.00	0.03	2.02
Bow_PubT_Conc_BibF	{antibodies,middle east}	{high}	67	1.00	0.03	2.02
Bow_PubT_Conc_BibF	{merscov,WoScategory_0_ aisQ_Q1_D1}	{high}	67	1.00	0.03	2.02
Bow_PubT_Conc_BibF	{flea,canine, FORD_0_ford_10600}	{low}	22	0.89	0.01	1.76
Bow_PubT_Conc_BibF	{coronavirus,isolate, FORD_0_impactQ_Q2}	{low}	22	0.89	0.01	1.76
ScispaCy	{simple antibody test methods}	{low}	689	0.58	0.53	1.1

Table 10 (continued)

Matrix	LHS	RHS	Supp.	Conf.	Cov.	Lift
PubTator_Conceptnet	{mers,Mice,us}	{high}	22	1.00	0.01	2.13
PubTator_Conceptnet	{mers,infection,Mice}	{high}	22	0.96	0.01	2.04
PubTator_Conceptnet	{canine,flea}	{low}	21	0.95	0.01	1.8
PubTator_Conceptnet	{mers,body,us}	{high}	44	0.94	0.02	2
PubTator_Conceptnet	{infection,infected}	{high}	22	0.65	0.02	1.39
PubTator	{recombinant fcov nucleocapsid protein rnp}	{low}	689	0.58	0.53	1.1
AuthorsNames	{woo,yuen kwok yung}	{high}	44	0.97	0.02	2.06
AuthorsNames	{patrick,yuen kwok yung}	{high}	22	0.97	0.01	2.05
AuthorsNames	{chan,patrick}	{high}	22	0.96	0.01	2.05
AuthorsNames	{chan,yuen kwok}	{high}	27	0.94	0.02	2.01
Bow	{canine,virus}	{low}	23	1.00	0.01	1.89
Bow	{merscov,mice}	{high}	28	1.00	0.01	2.13
Bow	{ifn,innate,respiratory}	{high}	22	1.00	0.01	2.13
Bow	{mice,protection,vaccine}	{high}	22	1.00	0.01	2.13
Bow	{hepatitis,study}	{low}	16	0.73	0.03	1.38
Bow	{associated,recently}	{high}	22	0.73	0.02	1.55

LHS antecedent of the rule, *RHS* prediction made by the rule, *Supp* number of articles matching the complete rule, *Conf* percentage of articles matching LHS for which the RHS is true (1 is 100%). *Cov* percentage of articles in the input dataset for which LHS is true, *Lift* is a ratio of the confidence of the rule (conf) and the expected confidence, which is the percentage of articles in the input dataset being assigned to the target class in the RHS of the rule

enriched (matrix PubTator_ConceptNet), the accuracy of a CORELS model consisting only of two rules matched the accuracy of a CBA model with 73 rules. For the remaining matrices, CBA outperformed CORELS, but the CBA models contained a noticeably larger number of rules. Whether this can be considered an accuracy-interpretability trade-off is unclear, since these rules provided possibly useful insights. Manually chosen representative selections from the rules generated by CBA for individual matrices are included in Table 10 and for CORELS in Table 11.

For a more in-depth analysis, we chose the matrix BOW_Pubtator_Conceptnet, since it contains most of the features related to article content and at the same time does not contain the bibliometric features, which are analyzed separately. The rule mining on this matrix also generated most rules; CBA returned 465 rules for the matrix version with additional feature cleaning. Figure 10 shows the clusters generated from these rules with setting $k = 30$, *lhs_items* = 3. Out of multiple evaluated configurations, this setting produced the best results according to our subjective evaluation.

When reviewing the rules, we noticed that multiple rules refer to specific animals (cf. Table 14 in the Appendix). These are visualized in the form of a graph in Fig. 11. These rules form two clusters. Rules referring to camels and bats (and their synonyms or word forms) are associated with a high citation count. In contrast, rules referring to other animals (dogs, cats, cows, rats, squirrels) are associated with a low citation count. A special case are mice, which are associated both with low and high citation counts. Rules in the CBA model are selected so that they are not redundant, however, the CBA rule list is a result of extensive rule pruning and therefore may not always be representative.

Table 11 Example rule lists generated by CORELS

Matrix	Rule list
ScispaCy	if [cos7 cells&& not wildtype di rna ne1 rna]: high_citation = True, else high_citation = False
PubTator_Conceptnet	if [not mers&& not body]: high_citation = False, else high_citation = True
ScispaCy_Conceptnet	if [chemoattractant&& not pegylated]: high_citation = True, else high_citation = False
PubTator	if [not human&& not MERS-CoV]: high_citation = False, else high_citation = True
AuthorsNames	if [not paul&& not peter]: high_citation = False, else high_citation = True
Bow	if [respiratory syndrome]: high_citation = True, else high_citation = False
Bow_ScispaCy	if [assessment&& not east]: high_citation = False, else high_citation = True
Bow_ScispaCy_Conceptnet	if [respiratory syndrome]: high_citation = True, else high_citation = False
Bow_PubTator	if [respiratory syndrome]: high_citation = True, else high_citation = False
Bow_PubTator_Conceptnet	if [respiratory syndrome&& not sars patients]: high_citation = True, else high_citation = False
BibliometricFeatures	if [not FORD_0_aisQ_Q1_D1&& not FORD_0_aisQ_Q1_D2]: high_citation = False, else high_citation = True
Bow_BibliometricFeatures	if [FORD_0_impactQ_Q2&& not middle east]: high_citation = False, else high_citation = True
Bow_PubT_Conc_BibFeatures	if [respiratory syndrome]: high_citation = True, else high_citation = False

Therefore, we analyzed all 178,098 candidate rules output by the apriori algorithm. The results shown in Table 12 confirm the same pattern that emerged from the clustering of the rules in the CBA model. For example, for 'camel', there were 503 rules in the high citation count category, but no rule predicting the low citation count. Conversely, rules referring to cats, dogs, or cows were always predicting the low citation category. Analysis of these reasons and the interpretation of this finding follows in the discussion section.

Analysis of author names

There are multiple rules associating first names with the high citation category. Table 15 presents a list of rules with the highest lift referring to author names. This includes both western-sounding names (Albert, Eric, Susanna, Christian) and, to a lesser degree, other names (Deng, Shibo). However, focusing on the rules predicting a low citation count, there

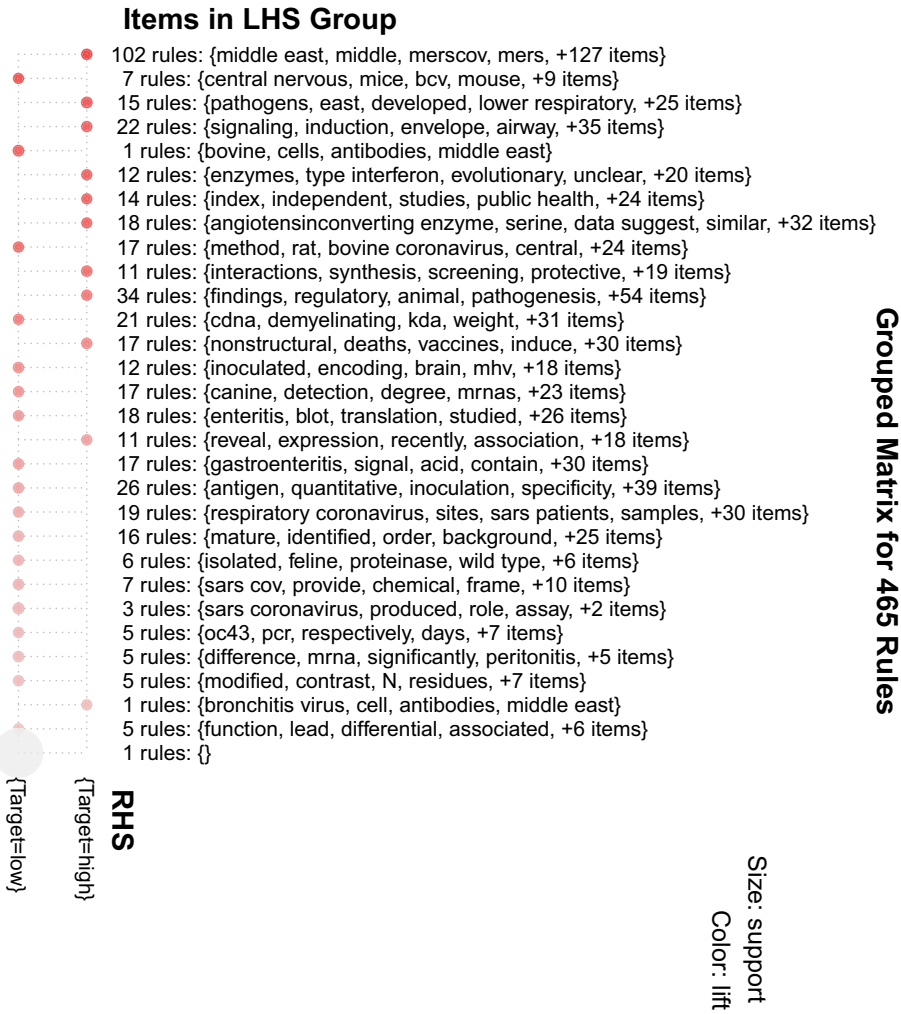

Fig. 10 Rule clustering results for CBA model generated on BOW_Pubtator_Conceptnet (version with additional cleaning)

is only one rule referring to a western sounding name (Nicola), but nearly 20 rules referring to non-western sounding names, and one rule referring to a combination.

We performed a deeper analysis of these names to clarify this phenomenon. As can be seen in Table 3, the names of the authors were processed using the BOW matrix, which has been reduced by removing words not occurring in at least 32 different documents ($min\_df = 32$). A total of 143 unigrams, 17 bigrams and 2 trigrams (considered as names) meeting the threshold were extracted from the author information. These names were assigned a nationality using the approach described by Ye et al. (2017)[8], which assigns multiple nationalities with different probabilities to each name. One of three continents (Africa, Asia, Europe) is consequently assigned based on the nationality. For our analysis,

[8] We used https://www.name-prism.com/, a system used by Ye et al. (2017) authors.

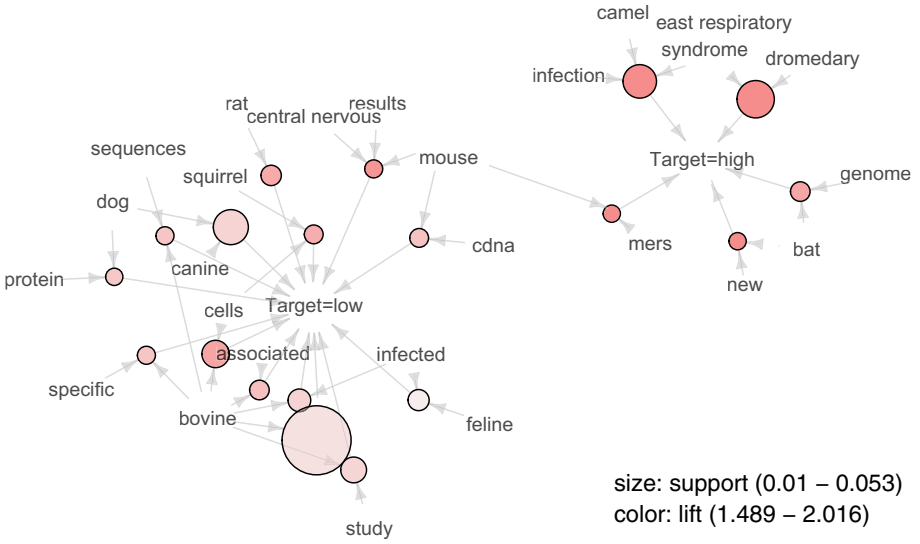

Fig. 11 Visualization of CBA rules related to animals from BOW_Pubtator_Conceptnet (version with extra feature cleaning). This graph was automatically generated by arulesViz (Hahsler & Karpienko, 2017), and subsequently edited for better readability (visually overlapping text and nodes were moved, no changes to the nodes, their labels, or their connections were made)

Table 12 Number of rules predicting the high/low categories containing the given concept in the antecedent

Concept	High	Low
Camel	503	0
Dromedary	575	0
Feline	0	162
Dog	0	35
Rat	0	6
Mouse	537	950
Bat	292	0
Cow	0	156
Bovine	0	200
Squirrel	0	8

The counts were generated from all candidate rules learned with the apriori algorithm from BOW_Pubtator_Conceptnet (version with extra feature cleaning)

we assigned the name to the continent associated with the highest probability. An overview of the results is in Table 13. Note that three names had the same probability for both continents, therefore they are counted twice in Table 13.

Table 13 The continent, nationality, and number of author names, based on (Ye et al., 2017)

Continent	Nationality	Number of names
Africa	Muslim-Nubian	2
Africa	African-WestAfrican	1
Africa	Muslim-Maghreb	1
Africa	African-EastAfrican	3
Asia	EastEasian-Malay-Indonesia	1
Asia	EastEasian-Malay-Malaysia	1
Asia	Muslim-ArabianPeninsula	1
Asia	EastAsian-Indochina-Myanmar	2
Asia	EastAsian-Chinese	64
Asia	EastAsian-South Korea	8
Asia	EastAsian-Indochina-Vietnam	9
Asia	EastAsian-Japan	4
Europe	Hispanic-Portuguese	3
Europe	Hispanic-Spanish	4
Europe	European-SouthSlavs	1
Europe	European-Italian-Romania	1
Europe	European-Italian-Italy	1
Europe	Europe- French	1
Europe	European-German	12
Europe	European-French	13
Europe	Celtic-English	31
Europe	Nordic-Finland	1
Total		165

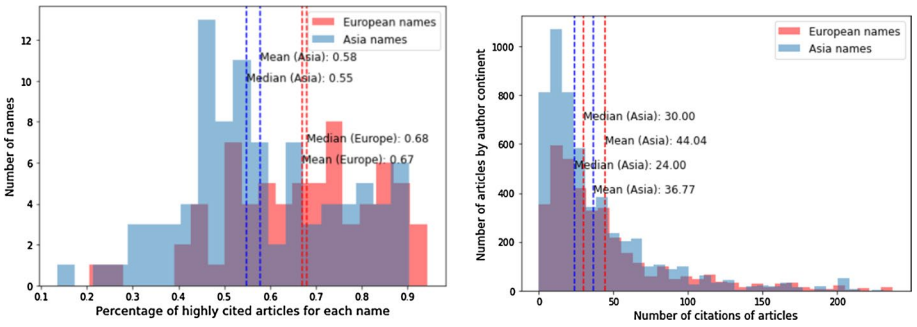

Fig. 12 Left side: distribution of the number of European and Asia names by the percentage of highly cited articles discretized by the number of bins = 20. Right side: Distribution of the number of articles with Europe and Asian author names by the number of citations

Individual names can be associated with different ratios of high citations. These ratios can range from 0 (all articles co-authored by given name are in the lowly cited category) to 1 (all articles are in the highly cited category).

Figure 12 (left) depicts these ratios and the corresponding count of names.[9] For example, author name 'Susanna' has in total 32 highly cited articles out of 36 total co-authored, which corresponds to ratio of 0.88 (plotted on the x-axis). Figure 12 (right) shows a different perspective. Authors with Asia-categorized names are more commonly co-authoring less cited articles than authors from Europe-categorized names.

We tested the statistical significance of the difference in the distributions shown in Fig. 12 (left). Based on the Mann-Whitney U test and the t-test, we can confirm at 1 percentage level of significance that the distributions are statistically significantly different. The same statistical test was applied on the distributions shown in Fig. 12 (right). There the distributions were also found to be statistically significantly different at 1 percentage level of significance. This provides further evidence for the results based on the interpretation of discovered rules.

Discussion

In this article, we have reported on an explorative analysis of factors influencing whether an academic paper will be cited or not. First, we analyze effects influencing predictive performance. The remainder of the discussion is structured by the group of features used.

Effect of dataset size The results have shown that accuracy of up to 83% can be obtained based on a relatively small training set. Increasing the training set size have not resulted in increased accuracy, but this can be partly attributed to the fact that the experiments with larger data were performed on a newer release of the underlying corpus, which was possibly harder as it had more articles from 2020. For these freshly published articles, fewer citations were available.

Regression vs classification task formulation Our results indicate that formulating the problem as a classification task already in the early stage of data processing leads to better results. One explanation is that the classification problem is more robust to noise because the number of citations that defines a given class of the target variable depends on the median of normalized (by age) number of citations. For example, even though we have shown that the used citation counts are in good agreement with high-quality WoS data, it could happen that citations for a specific paper are increased due to unhealthy citation practices. In the binary classification formulation, the same article would likely belong to the same class of the target variable even if these spurious citations have not been removed.

Author names There is a clear pattern showing that English-sounding author names are associated with a higher citation count. This observation is most succinctly characterized by the CORELS models for AuthorsNames featured in Table 11. After applying De Morgan's laws, a verbalized version of this classification model would read as:

If the list of author names includes **Paul** or **Peter**, the paper will be highly cited; otherwise, it will be lowly cited.

[9] Note that Africa was omitted due to paucity of data.

It should be noted that this classification model has the lowest accuracy of all models that we generated. Nevertheless, the underlying pattern is also confirmed by other models, including the explanation generated for a random forest model by LIME in Fig. 8.

What came as a surprise was that the CBA rule model learned on author names outperformed a rule model learned on the article content (AuthorsNames vs Bow in Table 7). Focusing on the random forest results in Table 6, the BOW model was slightly better, but we find it still remarkable that only author names are so predictive of citation count. Interestingly, using both pieces of information did not improve the performance (Bow_BibliometricFeatures in Table 6).

Bibliometric features The bibliometric features include journal Impact Factor (IF), Article Influence Score (AIS) and additional features derived from these measures. Since our analysis was built in a cumulative manner, the corresponding BibliometricFeatures matrix also contained the author names.

The results largely confirm the expectations. As Table 8 shows, considering bibliometric features only, AIS for the primary journal category has the highest feature importance, closely followed by the journal impact factor. English-sounding names were also included among the strongest predictors.

Impact factor and AIS were the strongest predictors overall, also considering all other predictors, including individual n-grams from the article content (Table 8—third column). Congruently with prior results (Vieira & Gomes, 2010), the author count came out as a modestly strong predictor. A possibly interesting implication from this analysis is that the Article Influence Score (AIS), the newer of the two measures of journal quality, does not substantially change citation prediction over the Journal Impact Factor. The latter is an older metric, which unlike AIS does not distinguish the quality of the citations. This result could be interpreted so that AIS serves as a comparably good predictor for the number of citations as the journal impact factor, while at the same time carrying the additional information on the impact of the journals in which the citations appeared.

Bag of words and entity enrichment As can be seen from the last pair of columns in Table 8, feature importance for n-grams computed for Random Forests does not provide much insight. Those found most important are very general n-grams (e.g., east respiratory syndrome) and are likely an artefact of the way the input corpus was constructed as articles to CORD-19 corpus were included based on their applicability to the research on Sars-Cov-2.

In contrast, the utility of using the extracted entities to represent article content can be seen on the second pair of columns in Table 8. In this case, the features correspond to entities. Out of the twenty top entities by feature importance (random forest), about ten correspond to subspecies of mammals (dromedary, camels, cow, rats, bovine, canine, human, humans, people, dog). A limitation of the interpretation of the Random Forest results is that the used feature importance score does not provide information on the direction, i.e., whether the presence of this entity is associated with a high/low citation count.

Rule learning results Figure 11 shows that the presence of words (or entities) referring to bats and camels is associated with high citation count, while other animals are associated with low citation count, except mice, which are associated with both categories. The relatively high number of rules referring to animals could be attributed to the fact that animals have been much debated in relation to coronavirus. One important context are the animal models, which were used for the study of viruses and therapeutics. Other contexts include the discussion of whether the animal can become infected, or is directly a source of infection, is a host for another virus which is a close known relative of a human coronavirus and can spread the infection to other host species. This is the case for dromedaries (or

camels) (Reusken et al., 2013; Sharun et al., 2020) and bats (MacFarlane & Rocha, 2020; Pereira et al., 2020; Poon et al., 2005), which are all predictive of the high citation category. Camels have been discussed as the source of infection, as have bats (Shereen et al., 2020). Camels are mainly associated with the MERS virus (Azhar et al., 2014), which is also confirmed by one of the rules. The combination of the words camel and human in one rule indicates that articles often discuss the transmission of the virus from camel to human. Examples of articles matching this rule include "Infection, Replication, and Transmission of Middle East Respiratory Syndrome Coronavirus in Alpacas" (Adney et al., 2016) or "Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a nationwide, cross-sectional, serological study" (Müller et al., 2015).

In contrast, most other animals, which have often been discussed in connection with coronavirus as possible spreaders (Muñoz-Fontela et al., 2020), are predictive of the low citation category. The same seems to also apply to dogs (also canines) and cats (or felines). Dogs and cats are common pets and many researchers investigated if they could suffer from coronavirus infections. However, until early 2020, there was essentially no evidence indicating that domestic animals like cats or dogs can be infected with Sars-Cov-2 (Goumenou et al., 2020).

Articles referring to mice were present in both categories. Mice are often used as model organisms for studying human diseases (Justice & Dhillon, 2016) and as such, are referred to from a wide variety of contexts. An example of rules representative of the different contexts (and citation categories) is given in the following two boxes. For each rule, we also provide example articles from CORD-19 covered by these rules.

Example rule: ***mouse,mers => Target=high, confidence=1.0, support=16.***
This rule covers 16 articles. All of them are correctly classified by the rule into the high category (confidence 100%). An example covered article (sample title):
- Alisporivir inhibits MERS- and SARS-coronavirus replication in cell culture, but not SARS-coronavirus infection in a mouse model

This article, published in 2017, has 12 citations, according to OpenCitations. After adjusting the citation count for the age of the article, we obtain $\frac{12}{2020-2017}=4$ normalized citations. The article thus falls into the high category according to Table 1, which contains articles with more than two citations per year.

Example rule: ***mouse,cdna => Target=low, confidence=0.9, support=18.***
Example covered articles (sample titles):
- A Murine and a Porcine Coronavirus Are Released from Opposite Surfaces of the Same Epithelial Cells
- Murine AKAP7 Has a 2',5'-Phosphodiesterase Domain That Can Complement an Inactive Murine Coronavirus ns2 Gene

The first article has 9 citations as retrieved from OpenCitations, and it was published in 1996. After the citation normalization, it has below 2 citations per year, falls into the low category, and is correctly classified as such by the rule.
The second article (*Murine AKAP7 ...*) has 17 citations, and it was published in 2014. After the citation normalization, it falls into the highly cited category. The rule *mouse,cdna => Target=low* covers this article, but incorrectly classifies it as low. However, the CBA classifier contains another matching higher priority rule, *interferon ifn,replication => Target=high, conf=1.0, supp=18*, which correctly classifies this article.

Note that the citation counts were retrieved at the time our article was under preparation and may have changed since then. Additionally, citation counts on other services like Google Scholar may be higher.

Biological Interpretation Based on a phylogeny of coronaviruses and their clades (Chan et al., 2015) we interpreted these results with respect to phylogenetic groupings and

distances. Bat coronaviruses are found in at least 4 different coronavirus clades, including the group betaB with human SARS, suggesting a high phylogenetic diversity of coronaviruses in bats. This is consistent with bats harboring persistent viral infections with enhanced viral shedding (Subudhi et al., 2019). It may also be that bats have been more studied in this respect, and we are observing a host and virus sequencing bias. Of note, a bat coronavirus is the only other virus from a non-human species in the betaB clade along with the SARS-CoV and SARS-CoV-2 viruses. The camel coronavirus (MERS) is in a sister betaC clade, also close to human SARS coronaviruses. Thus both species linked to high citation counts harbor coronaviruses which are more phylogenetically similar to human SARS viruses. On the other hand, feline (FIPV, FCOV) and canine coronaviruses (CCOV) are in the alpha coronavirus clade (Whittaker et al., 2018) and more distant from the betaB clade with human SARS viruses. There is also a murine coronavirus MHV, which is in the betaA clade, further from both the SARS and MERS clades. Together, these results highlight a pattern of virus phylogenetic distance to human SARS-CoV and SARS-CoV-2 related to high and low citation counts. This would be consistent with closer phylogenetic distance allowing better molecular inferences and transfer of information and virus study results from one species (e.g. bat) to human.

Rules also uncovered multiple patterns referring to specific types of biomedical entities, such as drugs or other compounds or manifestations of diseases. From the group of 465 rules selected by CBA from the association rule learning results on Bow_Pubtator_Conceptnet (with extra cleaning), we selected rules with the highest lift value: 102 rules had lift 2.015. All these rules predicted the high citation category. A lift of 2.0 indicates that articles containing this concept are about twice as likely to belong to the highly cited category than an average article in the training data. Among these rules, several rules referred to biomedical entities of potential interest (interferons, protease, spike protein, peptidase, dpp4). Dipeptidyl peptidase 4 (DPP4) is a MERS-CoV receptor (Li et al., 2020), which is, along with interferons and proteases, a possible critical determinant for MERS-CoV pathogenesis and transmission—both inter and intraspecies (Widagdo et al., 2019). Also, DPP4 was considered as a candidate binding target of SARS-CoV-2 spike protein (Li et al., 2020). Since DPP4 inhibitors are considered as possible therapeutic targets for Sars-Cov-2 (Strollo & Pozzilli, 2020), we consider DPP4 as a particularly interesting focus for a targeted future rule learning analysis.

Limitations

The work presented in this article is one of the first attempts to predict citations based on the contents of research articles using machine learning techniques and to explain the predictions. We tried to select a representative selection of machine learning and explanation methods, as well as a current dataset for which such analysis could be validated by domain experts. We acknowledge that our work suffers from multiple limitations, which we have explained and discussed further below, and some of which we would like to address in follow-up work.

The quality of citation data was also limiting. In future work, we will thus consider supplementing OpenCitations with a commercial bibliometric service. We found that for 60% of articles uncovered by OpenCitations, citation counts could be retrieved from Web of Science. The predictive performance could also possibly be improved by involving additional features, such as the country and reputation of the authors' institutions, and the h-index of authors, and features of citation, like how prominent the citation in the source paper is.

In our analysis, we have controlled for the age of the publication, but not for the rank of the journal in which it appeared. This could be done, e.g., by performing the analysis only for articles from journals assigned to the same quartile by impact factor or AIS. The combination of the most recent natural language processing techniques—BERT combined with a neural network classifier—yield the best performance. The accuracy of the BERT model would have likely further increased if we retrained the BERT model on research literature instead of using a generic model trained on Wikipedia. In addition, more thorough tuning of hyperparameters could have led to improved results. Both of these directions would, however, require considerable computational resources. Possibly the most promising approach that could also address the lower performance of the entity detection coupled with semantic enrichment from external sources is the utilization of graph-based embedding techniques, such as RDF2vec (Ristoski et al., 2019). In terms of interpretability, future work could employ some of the recently proposed modifications of LIME that aspire to address some of its shortcomings. A particularly promising direction can provide LioNets, which were shown to generate more precise explanations than LIME (Mollas et al., 2019). Another appealing direction is the use of explanation techniques combining relational rule learning and deep learning (Schmid & Finzel, 2020) techniques on a graph-based version of COVID-19 research data (Reese et al., 2020).

Conclusion

Research articles get cited for many different reasons. Most prior works focused on those attributable to general bibliometric factors, such as the quality of the journal, whether the article is openly available or not, and the number of authors. In this work, we have attempted to link research interest to the content of the article. Due to the paucity of prior work in this area, there were limited clues as to which group of methods would yield the best results on this problem. In our study, we have tried to address this research gap by applying a representative choice of preprocessing, data analysis, and interpretation techniques.

We were disappointed with the performance of enrichment with entity detection methods, which we hoped could improve results over the standard bag-of-words approach. Nevertheless, when combined with rule learning, the rules learned from the entity-based representation were subjectively more interpretable. Overall, our research confirmed the applicability of the interpretability-accuracy trade-off. The best predictive performance was obtained with a "black-box" method—neural network classifier over BERT-based text representation. The rule-based models yielded the most insights. In our work, we have shown how both techniques can be combined. We used random forests to evaluate data preprocessing setups and additionally used their results evaluated with eXplainable Artificial Intelligence (XAI) techniques to support a fine-grained interpretation based on rules.

One of the unique elements of our research is the rule-based approach, which provides both local classification and insight. Such rules can be useful also once a selection of a research topic and a method has been made. Our results provide information about animal species as virus hosts, which are of high interest in the context of COVID-19 research. Given a specific topic, the extracted rules can also provide more general guidance on which combination of a journal and a license for the distribution of the content, or a specific preprint server, has been associated with most citations.

Table 14 Example of the rules considering animals for CBA algorithm

Matrix	LHS	RHS	Support	Confidence	Lift
Bow_ScispaCy	{evolutionary flexibility,animal models}	{Target=high}	16	0.76	1.53
Bow_ScispaCy_Conceptnet	{theoretic,animal models}	{Target=high}	16	0.76	1.53
Bow_PubTator	{animal,evidence}	{Target=high}	18	0.90	1.81
Bow_PubTator_Conceptnet	{us,animal}	{Target=high}	25	1.00	2.01
Bow_BibliometricFeatures	{animals,zoonotic}	{Target=high}	18	1.00	2.01
Bow_BibliometricFeatures	{journal_Journal_of_ Virology,animals}	{Target=high}	30	0.85	1.72
Bow_Pubtator_Conceptnet _BibliometricFeatures	{cov,animal}	{Target=high}	25	1	2.01
Bow	{animals,middle east}	{Target=high}	26	1.00	2.01
Bow	{animals,zoonotic}	{Target=high}	18	1.00	2.01
Bow	{animal,evidence}	{Target=high}	18	0.90	1.81
Bow_ScispaCy	{dogs}	{Target=low}	22	0.79	1.55
Bow_BibliometricFeatures	{canine,dogs}	{Target=low}	21	0.88	1.73
Bow_PubTator_Conceptnet	{Cats,feline coronavirus,infectious_y}	{Target=low}	18	0.81	1.62
PubTator_Conceptnet	{coronavirus,cat,Cats}	{Target=low}	22	0.73	1.45
PubTator	{cats,feline coronavirus}	{Target=low}	20	0.71	1.41
Bow_ScispaCy	{bats rhinolophus ferrumequinum, demyelinating}	{Target=low}	16	1.00	1.98
Bow_ScispaCy_Conceptnet	{bats}	{Target=high}	47	0.80	1.60
Bow	{bats,coronavirus, east respiratory syndrome}	{Target=high}	21	1.00	2.01
Bow	{bats,coronavirus,transmission}	{Target=high}	17	1.00	2.01
Bow	{bats,respiratory,virus}	{Target=high}	22	0.96	1.92
PubTator	{rats}	{Target=low}	16	0.84	1.67
Bow_BibliometricFeatures	{camels,east respiratory syndrome, infection}	{Target=high}	31	1.00	2.01
Bow_BibliometricFeatures	{camel,east respiratory syndrome}	{Target=high}	21	1.00	2.01
PubTator_Conceptnet	{mers,camel}	{Target=high}	35	1.00	2.01
PubTator_Conceptnet	{humans,camel}	{Target=high}	25	1.00	2.01
PubTator_Conceptnet	{camel,coronavirus}	{Target=high}	16	1.00	2.01
PubTator_Conceptnet	{camel,infection}	{Target=high}	31	0.97	1.95
PubTator	{humans,camels}	{Target=high}	22	1.00	2.01
PubTator	{camel,camels}	{Target=high}	17	1.00	2.01
PubTator	{camel}	{Target=high}	24	0.96	1.93
PubTator	{camels}	{Target=high}	46	0.96	1.93
Bow_ScispaCy_Conceptnet	{dromedary}	{Target=high}	36	0.97	1.96
PubTator_Conceptnet	{mers,dromedary}	{Target=high}	26	1.00	2.01
PubTator	{humans,dromedary}	{Target=high}	18	1.00	2.01
Bow	{dromedary,east respiratory syndrome}	{Target=high}	35	1.00	2.01

Table 15 Example of the author names rules for CBA algorithm

LHS	RHS	Support	Lift
{m_ller marcel}	{Target=high}	0.012211	2.015544
{abdullah}	{Target=high}	0.010925	2.015544
{baric ralph,mark}	{Target=high}	0.010283	2.015544
{peter,van}	{Target=high}	0.013496	1.923928
{memish}	{Target=high}	0.012853	1.919566
{lai,michael}	{Target=low}	0.010925	1.874433
{wang,zheng}	{Target=high}	0.010925	1.903569
{yi,yuen kwok}	{Target=high}	0.010925	1.903569
{haagmans}	{Target=high}	0.015424	1.860502
{lai}	{Target=low}	0.017352	1.786224
{baker susan}	{Target=high}	0.010925	1.803382
{drosten christian}	{Target=high}	0.015424	1.791595
{kwok yung,woo}	{Target=high}	0.017995	1.763601
{woo patrick}	{Target=high}	0.017352	1.755474
{m_ller}	{Target=high}	0.012853	1.752647
{albert}	{Target=high}	0.012211	1.740697
{hung,woo}	{Target=high}	0.012211	1.740697
{woo,yuen}	{Target=high}	0.018638	1.719141
{jian,zheng}	{Target=high}	0.010925	1.713212
{huang,yi}	{Target=high}	0.010925	1.713212
{eric}	{Target=high}	0.025064	1.708831
{poon}	{Target=high}	0.014139	1.70546
{chan,patrick}	{Target=high}	0.014139	1.70546
{chan,yuen kwok}	{Target=high}	0.017352	1.700615
{peiris malik}	{Target=high}	0.010283	1.6973
{te,tseng}	{Target=high}	0.010283	1.6973
{li,zheng}	{Target=high}	0.010283	1.6973
{li,wang,yi}	{Target=high}	0.010283	1.6973
{berend}	{Target=high}	0.013496	1.693057
{baric}	{Target=high}	0.029563	1.685728
{woo}	{Target=high}	0.022494	1.67962
{wang,yi}	{Target=high}	0.012853	1.67962
{susanna}	{Target=high}	0.015424	1.668036
{christian}	{Target=high}	0.020566	1.65378
{deng}	{Target=high}	0.011568	1.649081
{fang,li}	{Target=high}	0.011568	1.649081
{graham}	{Target=high}	0.014139	1.642295
{lau}	{Target=high}	0.018638	1.623633
{patrick}	{Target=high}	0.023136	1.612435
{ali}	{Target=high}	0.015424	1.612435
{du,jiang}	{Target=high}	0.010283	1.612435
{li,zhou}	{Target=high}	0.010283	1.612435
{chan,yi}	{Target=high}	0.010283	1.612435
{stefan}	{Target=high}	0.012211	1.595639
{al}	{Target=high}	0.023136	1.577382
{baker}	{Target=high}	0.011568	1.577382

Table 15 (continued)

LHS	RHS	Support	Lift
{shibo}	{Target=high}	0.011568	1.577382
{te}	{Target=high}	0.016067	1.574644
{zheng}	{Target=high}	0.017995	1.567645
{matthew}	{Target=high}	0.01928	1.550418
{buonavoglia}	{Target=low}	0.012853	1.526688
{vincent}	{Target=high}	0.014781	1.54525
{van}	{Target=high}	0.041131	1.535653
{jan,peter}	{Target=high}	0.010283	1.535653
{peter}	{Target=high}	0.044344	1.52827
{mark}	{Target=high}	0.025064	1.511658
{li,yi}	{Target=high}	0.015424	1.511658
{christopher}	{Target=high}	0.012211	1.472898
{joo}	{Target=low}	0.012211	1.450353
{yee}	{Target=low}	0.012211	1.450353
{xiang}	{Target=high}	0.015424	1.46585
{alexander}	{Target=high}	0.010283	1.46585
{liu ding}	{Target=high}	0.010283	1.46585
{haan}	{Target=high}	0.010283	1.46585
{ching}	{Target=low}	0.010283	1.443414
{jiang,liu}	{Target=high}	0.010283	1.46585
{chen,wei}	{Target=low}	0.011568	1.42898
{jan}	{Target=high}	0.017995	1.447057
{andrew}	{Target=high}	0.021208	1.445934
{jian}	{Target=high}	0.017352	1.432097
{yuan}	{Target=low}	0.018638	1.403808
{peiris}	{Target=high}	0.015424	1.422737
{zhou}	{Target=high}	0.019923	1.420042
{nicola}	{Target=low}	0.012211	1.396636
{paul}	{Target=high}	0.037918	1.41568

The patterns obtained with the methodology described in our paper can be used to help shape research plans, as it provides a perspective on which concepts and their combinations resulted in follow-up scientific interest and which did not. Possible applications of our work thus include utilization in platforms for sharing and analyzing scientific articles. We envisage that obtaining insights on which combination of topics has a scientific impact, could be useful for research lab managers, and other decision makers in science, such as those allocating budget to individual research areas.

Appendix

See Tables 14 and 15.

Acknowledgements The authors would like to thank Vojtech Svatek for providing helpful feedback on an early version of the manuscript, and Karel Douda for acquiring citation counts for the V2 version of our dataset. The authors would also like to Clarivate—the producer of Web of Science and Journal Citation Reports—for providing an extra API call quota for access to WOS API Expanded.

Author contributions LB code, neural network, random forest and BERT models, experiments, editing. MPJ conception, interpretation, editing. TK conception, manuscript text, rule models, code, experiments, managed the research. GR code, initial manuscript, entity representation, experiments, literature review, editing the article. VS wrote part of the code for the bibliometric analysis and editing. All authors read and approved the final manuscript.

Funding GR was partly supported by grant IGA 43/2020 "Knowledge Engineering of Researcher Data (KNERD)". TK was supported by the Faculty of Informatics and Statistics, VSE, through long-term support for research activities and by a Horizon 2020 project HeartBIT_4.0 (grant id 857446). The citation data were acquired with support of CIMPLE project (CHIST-ERA-19-XAI-003). M.P.J. was supported by a grant from the Laboratory Directed Research and Development (LDRD) Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

Data availability URL links for datasets, where applicable, is provided in the article text. Results of processing are available from the authors at https://github.com/beranoval/why_is_cited.

Code availability The study has been conducted only with openly available software packages, which are referenced from article text along with the description of the salient settings. Scripts for the execution of the experiments are available from the authors at https://github.com/beranoval/why_is_cited

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical Standards Not applicable—the research did not involve human participants.

References

Adney, D. R., Bielefeldt-Ohmann, H., Hartwig, A. E., & Bowen, R. A. (2016). Infection, replication, and transmission of middle east respiratory syndrome coronavirus in alpacas. *Emerging Infectious Diseases, 22*(6), 1031.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research, 18*(1), 8753–8830.

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., & Madani, T. A. (2014). Evidence for camel-to-human transmission of MERS coronavirus. *New England Journal of Medicine, 370*(26), 2499–2505.

Belikov, A. V., & Belikov, V. V. (2015). A citation-based, author-and age-normalized, logarithmic index for evaluation of individual researchers independently of publication counts. *Research, 4,* 884.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cadorel, L., & Tettamanzi, A. G. B. (2020). Mining RDF data of COVID-19 scientific literature for interesting association rules. In *The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20)*.

Chan, J. F. W., Lau, S. K. P., To, K. K. W., Cheng, V. C. C., Woo, P. C. Y., & Yuen, K.-Y. (2015). Middle east respiratory syndrome coronavirus: Another zoonotic betacoronavirus causing SARS-like disease. *Clinical Microbiology Reviews, 28*(2), 465–522.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16,* 321–357.

de Winter, J. C. F. (2015). The relationship between tweets, citations, and article views for PLOS ONE articles. *Scientometrics, 102*(2), 1773–1779.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american*

chapter of the association for computational linguistics: Human language technologies (Long and Short Papers) (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science, 1*(2), 115–137. https://doi.org/10.1080/00033793600200111.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence, 33,* 3681–3688.

Giosa, D., & Di Caro, L. (2020) What2cite: Unveiling topics and citations dependencies for scientific literature exploration and recommendation. In *International conference on knowledge engineering and knowledge management* (pp. 147–157). Springer.

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics, 63*(1), 163–180.

Goumenou, M., Spandidos, D. A., & Tsatsakis, A. (2020). Possibility of transmission through dogs being a contributing factor to the extreme Covid-19 outbreak in North Italy. *Molecular Medicine Reports, 21*(6), 2293–2295.

Hahsler, M., Johnson, I., Kliegr, T., & Kucha, J. (2019). Associative classification in r: arc, arulesCBA, and rCBA. *R Journal, 9*(2), 254.

Hahsler, M., & Karpienko, R. (2017). Visualizing association rules in hierarchical groups. *Journal of Business Economics, 87*(3), 317–335.

Iqbal, F., Debbabi, M., & Fung, B. C. M. (2020). Authorship attribution using customized associative classification. In *Machine learning for authorship attribution and cyber forensics* (pp. 105–120). Springer.

Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258–263.

Justice, M. J., & Dhillon, P. (2016). Using the mouse to model human disease: Increasing validity and reproducibility.

Kaldas, M., Michael, S., Hanna, J., & Yousef, G. M. (2020). Journal impact factor: A bumpy ride in an open space. *Journal of Investigative Medicine, 68*(1), 83–87.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984–998.

Kliegr, T., & Kuchař, J. (2019). Tuning hyperparameters of classification based on associations (CBA). In *ITAT* (pp. 9–16).

Kuchař, J., & Kliegr, T. (2014). Bag-of-entities text representation for client-side (video) recommender systems. In *Proceedings of the RecSysTV*.

Kumar, M., Mazumder, P., Mohapatra, S., Thakur, A. K., Dhangar, K., Taki, K., et al. (2020). A chronicle of SARS-CoV-2: Seasonality, environmental fate, transport, inactivation, and antiviral drug resistance. *Journal of Hazardous Materials, 405,* 12–4043.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016) Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260–270). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1030

Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In: *NAACL 11 March 2016.* arXiv:abs/1603.03827

Li, Yu., Zhang, Z., Yang, L., Lian, X., Xie, Y., Li, S., et al. (2020). The mers-cov receptor dpp4 as a candidate binding target of the sars-cov-2 spike. *Iscience, 23*(6), 101160.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems, 26,* 431–439.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 2522–5839.

MacFarlane, D., & Rocha, R. (2020). Guidelines for communicating about bats to prevent persecution in the time of COVID-19. *Biological Conservation, 248,* 108650.

Mahmud, M., Kaiser, M. S., & Hussain, A. (2020). Deep learning in mining biological data. *arXiv preprint*arXiv:2003.00108

Mollas, I., Bassiliades, N., & Tsoumakas, G. (2019). Lionets: Local interpretation of neural networks through penultimate layer decoding. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 265–276). Springer.

Müller, M. A., Meyer, B., Corman, V. M., Al-Masri, M., Turkestani, A., Ritz, D., Sieberg, A., Aldabbagh, S., Bosch, B.-J., Lattwein, E., et al. (2015) Presence of middle east respiratory syndrome coronavirus antibodies in Saudi Arabia: A nationwide, cross-sectional, serological study. *The Lancet Infectious Diseases, 15*(5), 559–564.

Muñoz-Fontela, C., Dowling, W. E., Funnell, S. G. P., Gsell, P.-S., Riveros-Balta, A. X., Albrecht, R. A., et al. (2020). Animal models for COVID-19. *Nature, 586*(7830), 509–515.

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task* (pp. 319–327). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5034

Oermann, M. H., Nicoll, L. H., Ashton, K. S., Edie, A. H., Amarasekara, S., Chinn, P. L., et al. (2020). Analysis of citation patterns and impact of predatory sources in the nursing literature. *Journal of Nursing Scholarship, 52*(3), 311–319.

Pereira, M. J. R., Bernard, E., & Aguiar, L. (2020). Bats and COVID-19: villains or victims? *Biota Neotropica, 20*(3).

Piskorski, J., Haneczok, J., & Jacquet, G. (2020). New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? In *Proceedings of the 28th international conference on computational linguistics* (pp. 6663–6678).

Poon, L. L. M., Chu, D. K. W., Chan, K.-H., Wong, O. K., Ellis, T. M., Leung, Y. H. C., et al. (2005). Identification of a novel coronavirus in bats. *Journal of Virology, 79*(4), 2001–2009.

Ravanmehr, V., Blau, H., Cappelletti, L., Fontana, T., Carmody, L., Coleman, B., George, J., Reese, J., Joachimiak, M., Bocci, G., et al. (2021). Supervised learning with word embeddings derived from pubmed captures latent knowledge about protein kinases and cancer. *bioRxiv*.

Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., et al. (2020). Kg-covid-19: A framework to produce customized knowledge graphs for covid-19 response. *Patterns, 2*(1), 100155.

Reusken, C. B. E. M., Haagmans, B. L., Müller, M. A., Gutierrez, C., Godeke, G.-J., Meyer, B., et al. (2013). Middle east respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: A comparative serological study. *The Lancet Infectious Diseases, 13*(10), 859–866.

Rezaee-Zavareh, M. S. & Karimi-Sari, H. (2020). Effect of published papers by the institute for health metrics and evaluation on the impact factor of the lancet journal. *Journal of Investigative Medicine*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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* (pp. 1135–1144).

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., & Paulheim, H. (2019). Rdf2vec: Rdf graph embeddings and their applications. *Semantic Web, 10*(4), 721–752.

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design, 34,* 1013.

Roldan-Valadez, E., Orbe-Arteaga, U., & Rios, C. (2018). Eigenfactor score and alternative bibliometrics surpass the impact factor in a 2-years ahead annual-citation calculation: A linear mixed design model analysis of radiology, nuclear medicine and medical imaging journals. *La Radiologia Medica, 123*(7), 524–534.

Ruano, J., Aguilar-Luque, M., Gómez-Garcia, F., Alcalde Mellado, P., Gay-Mimbrera, J., Carmona-Fernandez, P. J., et al. (2018). The differential impact of scientific quality, bibliometric factors, and social media activity on the influence of systematic reviews and meta-analyses about psoriasis. *PLoS ONE, 13*(1), 191124.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Schmid, U., & Finzel, B. (2020). Mutual explanations for cooperative decision making in medicine. *KI-Künstliche Intelligenz, 34*(1–7), 2020.

Sharun, K., Tiwari, R., Patel, S. K., Karthik, K., Yatoo, M. I., Malik, Y. S., et al. (2020). Coronavirus disease 2019 (COVID-19) in domestic animals and wildlife: advances and prospects in the development of animal models for vaccine and therapeutic research. *Human Vaccines & Immunotherapeutics, 16,* 3043.

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research, 24,* 91.

Soares, J., Bazarian, F. K., Tavares, R. R., Denise, K., Bresciani, S., Pestana, R. C., et al. (2015). A review of the state of the art of self-citations. *Journal of Education & Social Policy*.
Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the thirty-first AAAI conference on artificial intelligence*, AAAI'17 (pp. 4444–4451). AAAI Press.
Strollo, R., & Pozzilli, P. (2020). Dpp4 inhibition: preventing sars-cov-2 infection and/or progression of covid-19? *Diabetes/Metabolism Research and Reviews, 36*(8), e3330.
Subudhi, S., Rapin, N., & Misra, V. (2019). Immune system modulation and viral persistence in bats: Understanding viral spillover. *Viruses, 11*(2), 192.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98.
Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics, 111*(2), 1053–1070.
Vieira, E. S., & Gomes, J. A. N. F. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4*(1), 1–13.
Wainberg, M., Alipanahi, B., & Frey, B. J. (2016). Are random forests truly the best classifiers? *The Journal of Machine Learning Research, 17*(1), 3837–3841.
Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al. (2020). Cord-19: The covid-19 open research dataset. *ArXiv*.
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology, 69*(2), 290–304.
Web of Science Group. Journal impact factor - journal citation reports. (2022). https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/
Wei, C.-H., Kao, H.-Y., & Zhiyong, L. (2013). Pubtator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Research, 41*(W1), W518–W522.
Whittaker, G. R., André, N. M., & Millet, J. K. (2018). Improving virus taxonomy by recontextualizing sequence-based classification with biologically relevant data: The case of the alphacoronavirus 1 species. *MSphere, 3*(1), e00463.
Widagdo, W., Ayudhya, S. S. N., Hundie, G. B., & Haagmans, B. L. (2019). Host determinants of mers-cov transmission and pathogenesis. *Viruses, 11*(3), 280.
Yamada, I., & Shindo, H. (2019). Neural attentive bag-of-entities model for text classification. *arXiv preprint*arXiv:1909.01259
Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S. (2017). Nationality classification using name embeddings. In *2017 ACM on Conference on Information and Knowledge Management*. arXiv:abs/1708.07903
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE international conference on computer vision (ICCV)*.

Authors and Affiliations

Lucie Beranová[1] · Marcin P. Joachimiak[2] · Tomáš Kliegr[3] · Gollam Rabby[3] · Vilém Sklenák[4,5]

✉ Tomáš Kliegr
tomas.kliegr@vse.cz

[1] Department of Econometrics, Faculty of Informatics and Statistics, VSE Praha, W Churchill sq. 4, Prague, Czech Republic

[2] Environmental Genomics and Systems Biology Division at Lawrence Berkeley National Laboratory, Berkeley, USA

[3] Department of Information and Knowledge Engineering, VSE Praha, Prague, Czech Republic

[4] Centre of Information and Library Services, VSE Praha, Prague, Czech Republic

[5] Department of Information and Knowledge Engineering, VSE Praha, Prague, Czech Republic