Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Medical Informatics 141 (2020) 104234

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Developing an automated mechanism to identify medical articles from wikipedia for knowledge extraction

Lishan Yu[a], Sheng Yu[b,c,d,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing, China
[b] Center for Statistical Science, Tsinghua University, Beijing, China
[c] Department of Industrial Engineering, Tsinghua University, Beijing, China
[d] Institute for Data Science, Tsinghua University, Beijing, China

ARTICLE INFO

Keywords:
Wikipedia
Text classification
Crawling classification
False discovery control

ABSTRACT

Wikipedia contains rich biomedical information that can support medical informatics studies and applications. Identifying the subset of medical articles of Wikipedia has many benefits, such as facilitating medical knowledge extraction, serving as a corpus for language modeling, or simply making the size of data easy to work with. However, due to the extremely low prevalence of medical articles in the entire Wikipedia, articles identified by generic text classifiers would be bloated by irrelevant pages. To control the false discovery rate while maintaining a high recall, we developed a mechanism that leverages the rich page elements and the connected nature of Wikipedia and uses a crawling classification strategy to achieve accurate classification. Structured assertional knowledge in Infoboxes and Wikidata items associated with the identified medical articles were also extracted. This automatic mechanism is aimed to run periodically to update the results and share them with the informatics community.

1. Introduction

Wikipedia contains rich biomedical information and has been widely used for medical informatics research [1]. In addition to basic text mining [2,3,4], Wikipedia articles can be also used for formal knowledge extraction. For example, the article titles, text written in bold, and redirections are usually medical concepts or named entities. The Infobox (the information box at the top right corner of each article), tables in the main text, and the Wikidata item associated with each Wikipedia article provides concept relations [5–7]. These medical concepts and relation can also be discovered from the free text, which are important research topics in natural language processing (NLP) [8–10]. These concepts and relations can be used to develop medical knowledge graphs that can provide high-level support to healthcare artificial intelligence [11,12], such as language understanding and decision support. In addition, the medical articles as a corpus can be used for training word/concept representations [13,14] and language models [15,16] to improve the modeling performance in various machine learning tasks. Therefore, although there are controversies about the scientific rigor and quality of some of the articles on Wikipedia [17–21], the size and richness of Wikipedia still make it one of the most useful data sources for medical informatics studies.

However, the size of Wikipedia also creates problems. Wikipedia is freely editable by internet users around the world, on any possible subject. As a result, medical articles only represent a tiny fraction of the entire Wikipedia. For example, the 2020-05-01 dump of Wikipedia contains over 20 million articles, which includes 14 million redirect pages and 6 million non-redirect articles. Among which, as our results indicate, about only 90 thousand articles (1.5 % of non-redirect pages, 0.5 % of all pages) are related to medicine. With such a tiny representation, using the entire Wikipedia for medical research can have negative effects. For example, language models trained on general text are less accurate in healthcare NLP than those trained with medical corpora [13,15,16], and medical term discovery and relation extraction models can have many false discoveries when applied to articles unrelated to medicine. In addition, the 2020-05-01 dump of Wikipedia is 65 GB in volume, and the 2020-06-01 dump of Wikidata is 1.1 TB when uncompressed, which creates unnecessary computational difficulties for researchers who only need the medical parts of them.

The goal of our work is to develop an automated mechanism to identify the medical article subset of Wikipedia, which can be used to facilitate further medical informatics studies. Currently, we look for articles on 7 categories of medical subjects: Anatomy (ANAT), Chemicals & Drugs (CHEM), Devices (DEVI), Disorders (DISO), Living

* Corresponding author at: Weiqinglou Rm 209, Center for Statistical Science, Tsinghua University, Beijing, 100084, China.
E-mail address: syu@tsinghua.edu.cn (S. Yu).

https://doi.org/10.1016/j.ijmedinf.2020.104234
Received 16 February 2020; Received in revised form 1 July 2020; Accepted 11 July 2020
Available online 13 July 2020

Beings (LIVB), Physiology (PHYS), and Procedures (PROC). The exact scope of these categories follows the semantic group definition of the Unified Medical Language System (UMLS) [22], with certain exclusions, as detailed in Supplementary Materials S1. For instance, for LIVB, we only included the semantic types of Bacterium, Fungus, Virus, and Eukaryote, which are more related to diseases than other live beings. Since there already exist multiple ontologies for genetics [23,24], we decided to exclude genetic concepts from our current search scope.

Semantic web projects and efforts associated with Wikipedia can be used to identify some of these categories [2,6,8,25]. For example, DBpedia [26] provides class labels that can help identify articles of certain categories, such as diseases and live beings, but it does not cover all target categories. Besides, DBpedia is not 100 % correct, and it updates slowly. Similarly, WikiProject Medicine also provides tags for several but not all interested semantic groups [27,28]. Therefore, instead of relying on existing semantic resources, we develop machine learning algorithms to identify medical articles and classify them into the aforementioned 7 semantic groups. As a side product, we also extract structured assertional knowledge from the Infoboxes and Wikidata items of these articles. As Wikipedia is constantly being updated by users, the automated mechanism allows us to periodically rerun to update our results and share them with the medical informatics community (https://github.com/yusir521/WikiMedSubset).

Identifying medical articles from Wikipedia and classifying them by semantic group pose a few uncommon challenges. The first challenge is the extremely low prevalence of each class. Generic text classification techniques have progressed rapidly in recent years, with some latest deep learning models exhibiting near-human accuracy [29–31]. Techniques have also been proposed to alleviate the sample imbalance issue [32–34]. However, Wikipedia articles are not plain text, but they have very rich elements and structures. To exploit these features to improve classification accuracy and efficiency, we devised a crawling classification strategy that only needs to classify a portion of Wikipedia articles, which can raise the prevalence and control the false discovery rate. We also incorporate various elements of Wikipedia pages into our models with feature engineering.

Another challenge to our work is acquiring annotated samples for training and validation. With the extremely low prevalence of each semantic category, manual annotation of a random sample is infeasible. For example, retrospectively estimating from our result, to acquire 50 sample articles on medical devices, it would require annotating 200 thousand non-redirect pages on average. To acquire sufficient annotated data, we employed the weak/distant supervision technique [35–37] and used the UMLS for automatic annotation. We also conducted limited manual validation on the model predicted medical articles.

The structure of the remaining paper is as follows. Section 2 gives an overall summary of the Wikipedia data and explains the preparation of the training data. Section 3 introduces the crawling classification strategy and models. Section 4 introduces baseline models for comparison and evaluation metrics. Section 5 shows the statistics of the identified articles and comparisons of model accuracy. Section 6 discusses various aspects of the results and compares the identified articles and extracted relations with possible alternative approaches. The last section summarizes the work and its limitations.

2. Materials and data preparation

Wikipedia is a website that is constantly being updated. The contents of Wikipedia are also available as dumps, which are backups of the website's database. The dumps are created every few months and are available for download. For this paper, we used the 2020-05-01 dump of Wikipedia. This dump contains 20,208,017 Wikipedia pages, among which 6,069,466 are non-redirect, i.e., they are actual articles.

We used the UMLS to create automatic annotations for training and validation. Non-redirect/disambiguation article titles were matched with the UMLS for concept recognition. To avoid ambiguities, we only used full string matches with UMLS "preferred terms", and terms with multiple possible concept matchings were abandoned. Eventually, 40,856 were identifiable as UMLS concepts. Among them, 11,843 articles/concepts were not in the chosen 7 semantic groups and were labeled as NULL. The composition of the matched articles is shown in Table 1. In the crawling classification, articles of the 7 target semantic groups are considered as positive samples, and the NULL class is considered as negative samples. In addition, given the extremely low prevalence of medical articles in the entire Wikipedia, we used a random sample of 17,000 Wikipedia articles whose titles could not be matched using the UMLS as additional negative samples; these samples were representative of the more unrelated articles. Therefore, the total automatic annotated samples had 29,013 positive and 28,843 negative. 80 % of these samples were used for training, and 20 % were used for testing.

Table 1
Composition of UMLS matched articles.

SemGroup	ANAT	CHEM	DEVI	DISO	LIVB	PHYS	PROC	NULL
Count	3,111	10,849	343	6,799	5,805	817	1,289	11,843

The crawling classification strategy, introduced in the next section, applies a breadth-first search to the Wikipedia articles. The breadth-first search requires at least one medical article (seed) in the search queue as a starting point. Indeed, since medical articles on Wikipedia are not guaranteed to be all connected (accessible from a sequence of links from any given medical article), it is necessary to use many articles as seed points to minimize the possibility of isolation. To find a large number of articles as seeds, we used Wikipedia's category hierarchy. A Wikipedia article is usually tagged with categories that are displayed at the bottom of the page (Fig. 1). The categories have a hierarchy: under each category, there can be subcategories as well as articles tagged with this category. We used articles within 5 steps down the Medicine and Anatomy categories to populate the search queue. These articles were likely to be in the defined scope of medical articles, and they were classified in the same way as other articles during the search. Additionally, UMLS-recognizable articles in the training set were also added to the seed list. The seed list eventually contained 225,239 articles.

3. Classification strategy and models

Our mechanism uses a two-step workflow, illustrated in Fig. 2: the first step identifies the medical subset of Wikipedia, and the second step classifies the articles (which were generally about medical concepts) by semantic group.

3.1. Step 1: the crawling classifier

To raise the prevalence of medical articles, the first step uses a crawling strategy. The crawler starts with a search queue filled with seed articles introduced in Section 2. At each step, the crawler uses a support vector machine (SVM) binary classifier to classify if the article is about medicine. If it is, links in the article to other Wikipedia articles will be extracted, and the linked articles will be added to the end of the queue to be classified, using the breadth-first search strategy; otherwise, the article will be abandoned and no linked articles will be added to the search queue. This crawling strategy leverages the fact that articles linked from a medical article are likely about medicine as well, so the process blocks the majority of the non-medical articles from being classified and keeps the positive rate high.

The SVM classifier uses the Gaussian kernel with the three kinds of features: (1) Naïve Bayes probabilities. We fit 4 Naive Bayes classifiers using word tokens from the main body, the section titles, the Infobox,

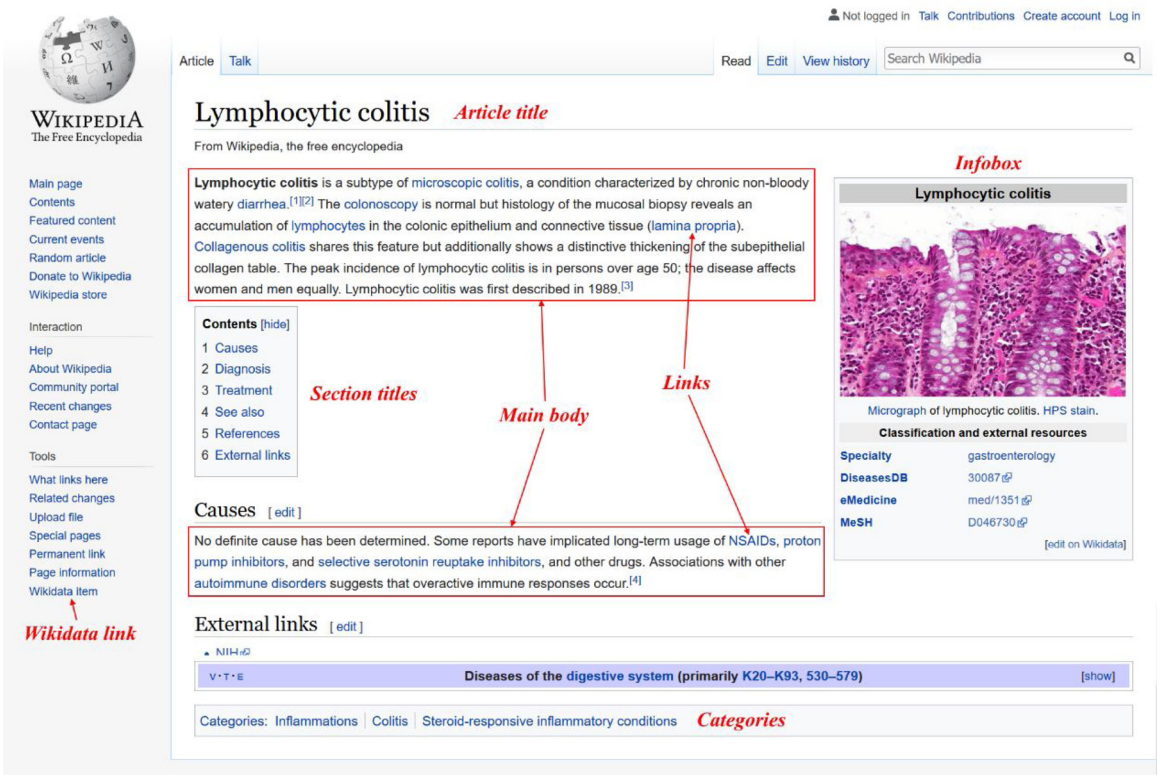

Fig. 1. Elements of a Wikipedia article (title, main body, Infobox, section titles, categories, and links).

and the categories, respectively (see Fig. 1 for illustration). Words from these fields usually exhibit a clear pattern that can help distinguish articles of different topics. The predicted probabilities from the classifiers that the article is about medicine are used as features, denoted by $x_{NB} \in \mathbb{R}^4$. Since each probability can be used for classification by itself, these 4 features are all strong predictors. (2) Article embedding. We use the Skip-gram model [38] to obtain 300-dimensional vector representations of stemmed words using the entire Wikipedia. Semantically close words are expected to have similar vector representations, thus words related to medicine are expected to cluster together. The embedding for an article, denoted by $x_{emb} \in \mathbb{R}^{300}$, is created by combining the vectors of the words in the main body with Term Frequency - Inverse Document Frequency (TF-IDF) weights. (3) Named entity recognition (NER). Three additional features, denoted by

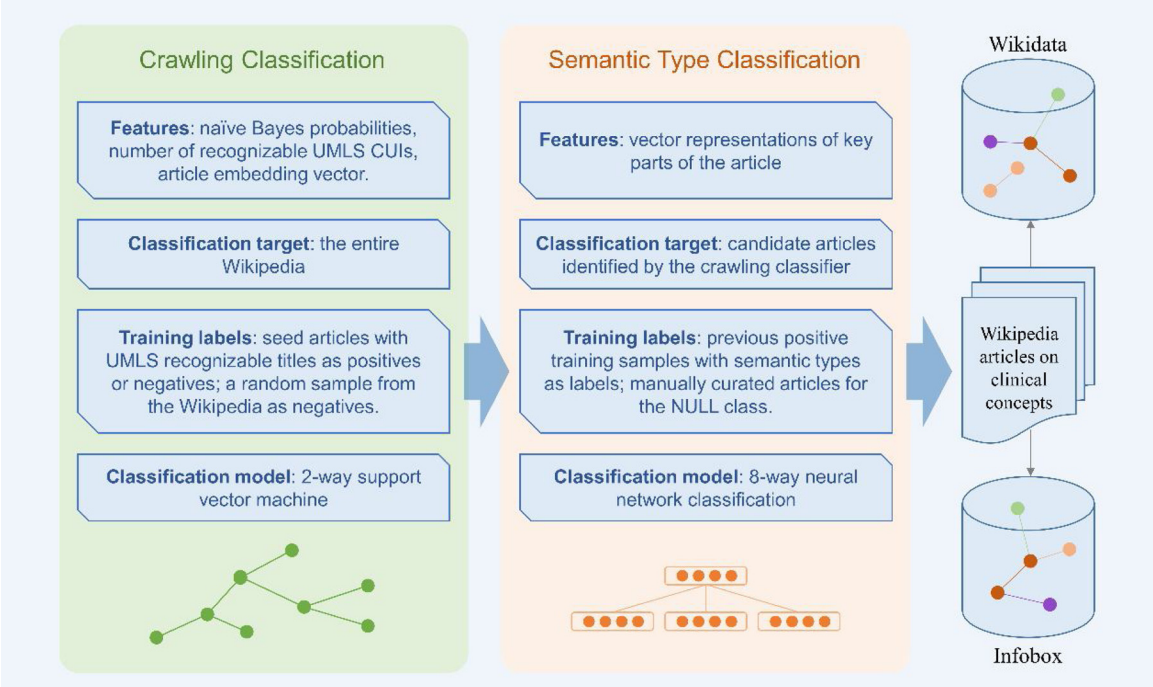

Fig. 2. The two-step classification workflow.

Table 2
The number of medical articles identified by the proposed mechanism (Proposed), NaïveB, RM-TF-IDF, and TextCNN.

	ANAT	*CHEM*	*DEVI*	*DISO*	*LIVB*	*PHYS*	*PROC*	*TOTAL*
Proposed	6863	35026	1502	14145	28524	2948	4412	93420
NaïveB	12544	62524	18764	18191	261697	16680	15068	405468
RM-TF-IDF	9058	46293	1911	18274	40841	2899	4610	123886
TextCNN	10719	55095	1806	33586	52317	4909	4862	163294

$x_{NER} \in \mathbb{R}^3$, are based on named NER with the software NILE [39] using the UMLS Metathesaurus as the dictionary. We use one binary feature to indicate if there is a recognizable concept within the target semantic groups after cue words, such as "is" and "are", in the first sentence of the article. Another two features count the number of recognized UMLS concepts in the whole article that are of and not of the target semantic groups, respectively. The features are combined as $x = (x_{NB}, x_{emb}, x_{NER}) \in \mathbb{R}^{307}$ as the input feature for the SVM, which is trained using the automatically labeled training set. Model parameters are tuned using 5-fold cross-validation.

3.2. Step 2: the semantic group classifier

Wikipedia articles classified as positive by the crawling classifier are further classified by a deep learning model to determine its semantic group. The classification is 8-way: the 7 target semantic groups and the NULL class, which is still present in the articles that are considered positive by the first classifier. To reflect the fact that the articles to be classified in the second step are much closer to medicine than those in the first step, the training data for the RNN only includes UMLS-matched articles and does not include the 17,000 randomly sampled articles.

After the crawling classifier in Step 1 has removed most of the non-medical pages, the category distribution should become sufficiently even that most mature text classifiers should work well. Our model uses three Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRU) [40] to generate vector representations for three parts of an article: the first sentence of the first paragraph (which usually states the nature of the topic entity), the remaining sentences of the first paragraph, and the first-level section titles. The three GRUs summarize the three text pieces to three vectors $x_{1st\_sen} \in \mathbb{R}^{32}$, $x_{1st\_para} \in \mathbb{R}^{64}$, and $x_{sec\_ttl} \in \mathbb{R}^{128}$, respectively. The three vectors are concatenated to a single vector $x_{concat} = (x_{1st\_sen}, x_{1st\_para}, x_{sec\_ttl})$, which is passed through a fully-connected layer with 150 neurons, then into the final layer for 8-way classification. The number of training epochs is determined by training the model on 84 % of the training data and validating on the reserved training data. The model is then retrained on the full training data.

4. Evaluation methods

We compared our mechanism with three off-the-shelf text classifiers: Naïve Bayes (NaïveB) [41], the logistic regression model with TF-IDF features (RM-TF-IDF) [42], and TextCNN [43], which usually attain excellent performance in semantic classification. All three models were trained on automatically labeled training data, including the randomly sampled Wikipedia pages labeled as NULL. TextCNN used kernel sizes 3, 4, and 5, with 100 channels. The embedding dimension for each word was 128. The baseline classifiers were applied to the entire Wikipedia for 8-way classification.

Two ways are used to evaluate the results from the proposed mechanism and the baseline models. The first way uses the 20 % automatically annotated samples reserved for testing, containing 11,571 samples. Recall, precision, and F score are calculated for each category. The second way randomly samples 100 articles predicted as medical articles from the result of each model and manually labels their categories (7 medical semantic groups + NULL). Accuracy and false discovery rate (the rate of NULL among articles predicted as medical) are calculated for each model.

We also considered and compared with alternative ways not using machine learning to identify Wikipedia medical articles. One such way is via Wikidata. A Wikidata item associated with a Wikipedia medical article may contain concept IDs from notable medical ontologies. Therefore, querying Wikidata items with such IDs can be used to identify medical articles in Wikipedia. We queried the 2020-06-01 dump of Wikidata for items that contained a concept ID from UMLS, RxNorm, NDF-RT, ICD-9, ICD-10, or LOINC to search for Wikipedia medical articles. We also compared our result with the 2020-06-01 version of DISNET [25], which was based on DBpedia and focused on diseases.

5. Results

The crawling classifier reached and classified 1,205,568 articles, which is 19.9 % of all the non-redirect Wikipedia articles. 111,900 articles were considered positive and were further classified by the second 8-way classifier, and 93,420 of them were classified into 1 of the 7 target medical semantic groups. In comparison, the baseline models NaïveB, RM-TF-IDF, and TextCNN predicted 405,468, 123,886, and 163,294 articles as medical articles, respectively. Table 2 shows the

Table 3
Precision (P), recall (R), and F-score (F) evaluated using the reversed articles with automatic labels.

	ANAT	*CHEM*	*DEVI*	*DISO*	*LIVB*	*NULL*	*PHYS*	*PROC*
Proposed P	94.44%	95.64%	83.33%	94.50%	96.85%	88.22%	60.53 %	80.90 %
NaïveB P	77.16 %	83.18 %	31.18 %	82.60 %	70.69 %	94.70 %	24.69 %	45.36 %
RM-TF-IDF P	92.24 %	91.92 %	69.57 %	90.92 %	94.15 %	91.28 %	48.00 %	76.47 %
TextCNN P	93.62 %	90.52 %	64.29 %	87.51 %	93.33 %	90.66 %	35.43 %	66.48 %
Proposed R	90.63 %	92.16 %	49.30 %	86.62 %	71.59 %	97.62 %	46.62 %	64.92 %
NaïveB R	86.75 %	91.60 %	40.85 %	77.68 %	95.16 %	82.18 %	40.54 %	53.23 %
RM-TF-IDF R	86.43 %	91.88 %	22.54 %	85.02 %	91.71 %	95.86 %	32.43 %	62.90 %
TextCNN R	80.61 %	91.32 %	12.68 %	83.03 %	91.80 %	95.29 %	30.41 %	47.98 %
Proposed F	92.50 %	93.87 %	**61.**95 %	90.39 %	82.32 %	92.68 %	52.67 %	72.04 %
NaïveB F	81.67 %	87.19 %	35.37 %	80.06 %	81.12 %	88.00 %	30.69 %	48.98 %
RM-TF-IDF F	89.24 %	91.90 %	34.04 %	87.87 %	92.91 %	93.52 %	38.71 %	69.03 %
TextCNN F	86.63 %	90.92 %	21.18 %	85.21 %	92.56 %	92.91 %	32.73 %	55.74 %

decomposition of the identified articles by semantic groups.

The baseline models predicted much more medical articles than the proposed method, which is undesirable because they were inaccurate and included many false discoveries. Table 3 shows the precision, recall, and F score evaluated using the reserved articles with automatic labels. The proposed mechanism achieved the best performance on almost every metric. It achieved the worst performance on NULL precision, that is, articles were incorrectly classified as non-medical. Among the 5,876 positive samples in the test set, 766 were misclassified as NULL, and 607 of which were misclassified by the crawling classifier. Interestingly, about half of the 607 were LIVB, most of which were not reached by the crawler, which is also reflected by the low recall of the category. This suggests that many UMLS-recognizable LIVB (viruses, bacteria, fungi, etc.) may not be linked to medical pages. And this may not be a drawback as it appears, as many microbes are not directly related to human health and thus are not desired in the medical subset. On the other hand, the proposed mechanism achieved the highest recall on the NULL category, which means that its results contain the fewest false discoveries. A high recall on NULL is an important property to have because most Wikipedia articles are non-medical, which means that a small drop on NULL recall would result in many false discoveries, as shown in Table 2. Table 4 further confirms this point. Based on manual review of the identified articles, the proposed mechanism has far higher positive sample accuracy than the baseline models, and it has the fewest false discoveries ("Richard Shope", "Isturgia", "Epichlorops", and "List of virus species" classified as LIVB, "Chlamys" and "Kiss curl" classified as DISO, "Hair-cutting shears" classified as ANAT, and "Pentamerida" classified as CHEM). Indeed, the false discovery rates of the baseline models are so high that their results are hardly usable, even though they are excellent text classifiers in general.

Combining our automatic search mechanism and medical ontology code queries, 110,850 Wikidata items in total can be found, as Fig. 3 shows. Among them, 91,513 can be identified by our mechanism, and 79,714 are exclusively identified from Wikipedia, showing our work is not replaceable by simple queries. This also partially suggests that our search has a high recall of medical articles on Wikipedia. Note that not all Wikidata items have associated Wikipedia pages. There are 87 million Wikidata items in the 2020-06-01 dump, but only 6 million Wikipedia non-redirect pages.

In comparing with DISNET, we found that of its 7324 diseases, 6979 (95.3 %) were classified as DISO by our search mechanism. There are 159 articles labeled as diseases by DISNET and were classified into other categories by our method, and many of our classifications were correct. There are also 186 articles labeled as diseases by DISNET but were not found by our mechanism. Many of these articles are about medicine and are true misses by our method, though they are not all diseases. Overall, judging from DISNET, our method's recall of diseases is very high. Finally, there are 7166 articles classified as DISO by our method but are not in DISNET, and they are generally correctly classified. Note that the UMLS DISO semantic group includes not only diseases but also other concepts, such as findings. However, many of the 7166 articles are indeed diseases. This shows that the DBpedia-based approach can still miss many pages. Samples of these comparisons are provided in Supplementary Materials S2-5.

Table 4
Accuracy of positive predictions, false discovery rate, the estimated number of medical articles with correct classification, and the estimated number of identified articles with wrong classifications by the proposed mechanism and the baselines.

	Positive accuracy	*False discovery rate*	*Est. medical articles*	*Est. wrong articles*
Proposed	**0.92**	**0.08**	85,946	7,474
NaïveB	0.26	0.72	**105,421**	300,046
RM-TF-IDF	0.69	0.26	85,481	38,405
TextCNN	0.53	0.39	86,546	76,748

6. Discussions

An automatic mechanism to periodically identify medical articles in Wikipedia and extract their structured knowledge is important to keep our medical informatics infrastructures up to date. For instance, "Coronavirus disease 2019" is already in our identified medical subset (the 2020-05-01 dump of Wikipedia), while it is not in DISNET (2020-06-01), which is DBpedia-based that is updated in a long cycle.

As discussed at the beginning, the major difficulty for developing a text classifier for the automatic mechanism is the extremely low prevalence of medical articles in Wikipedia. A high proportion of negative samples means a high false discovery rate for machine learning algorithms, which can potentially render the results useless. Therefore, the main goal of our design decisions is how to achieve a low false discovery rate while maintaining a high recall for medical articles. Instead of seeking more sophisticated deep learning text classification models, we decided to leverage the rich page elements and the connected nature of Wikipedia and developed the crawling classification strategy. The estimated numbers of identified medical articles in Table 4 show that our unique search mechanism did not sacrifice recall (compared to RM-TF-IDF and TextCNN, the two better models of the baselines), and its number of false discoveries is 1–2 levels of magnitude fewer than the baselines. In semantic group classification, as shown in Table 3, which is evaluated using the automatically annotated samples, our method still shines in most categories. The low recall of LIVB in Table 3 was because many pages of microbes (especially those not related to human health) were not connected with medical articles and they were not reached by the crawler. We do not consider this an issue at the moment until we can find better labels to differentiate microbes related to human health from those that are not.

To avoid missing medical articles, we used over 225 thousand seed articles in the breadth-first search, and the crawler eventually covered 20 % of Wikipedia articles, which we think is sufficiently large. Further raising the coverage would risk more false discoveries. We reviewed incorrect classifications by our method and found that many errors were due to articles being too short. For instance, the Wikipedia article 'Ancylobacter rudongensis' contained only one sentence: 'Ancylobacter rudongensis is a bacterium from the family of Xanthobacteraceae which has been isolated from root of the plant Spartina anglica from the beach from the Jiangsu Province in China' and was classified as 'NULL'. Short texts were particularly common in LIVB and caused many of our misclassifications. To accommodate the cases of insufficient information, our models avoided relying on a single source, such as the main text, and used various components of Wikipedia pages as features. For this reason, we did not conduct an ablation test and wanted to have feature redundancy. However, it appears that our models still need to improve on classifying very short articles. In terms of speed, although the crawling classification method only needed to classify 20 % of the non-redirect pages, it currently does not have a clear advantage in speed due to limited database optimization (crawling requires random reads of the database, while algorithms applied to the whole database use sequential reads). However, the time spent on training and applying the algorithms is little compared to the time spent on preprocessing Wikipedia and Wikidata (the 2020−06-01 dump of Wikidata is over 1TB in size and takes more than 1 day to decompress), so improving the speed is not of top priority.

Another difficulty and a major limitation to our study is the lack of gold-standard labels. As explained in Section 1, unbiased manual annotation is infeasible given the rareness of medical articles. Therefore, we used the UMLS for automatic labeling. The benefit of using the UMLS is that the generated sample size is very large. On the other hand, UMLS can introduce biased sample distribution to both training and validation. The labels are also imperfect. For example, we only want LIVB and CHEM that are related to human health, but UMLS cannot

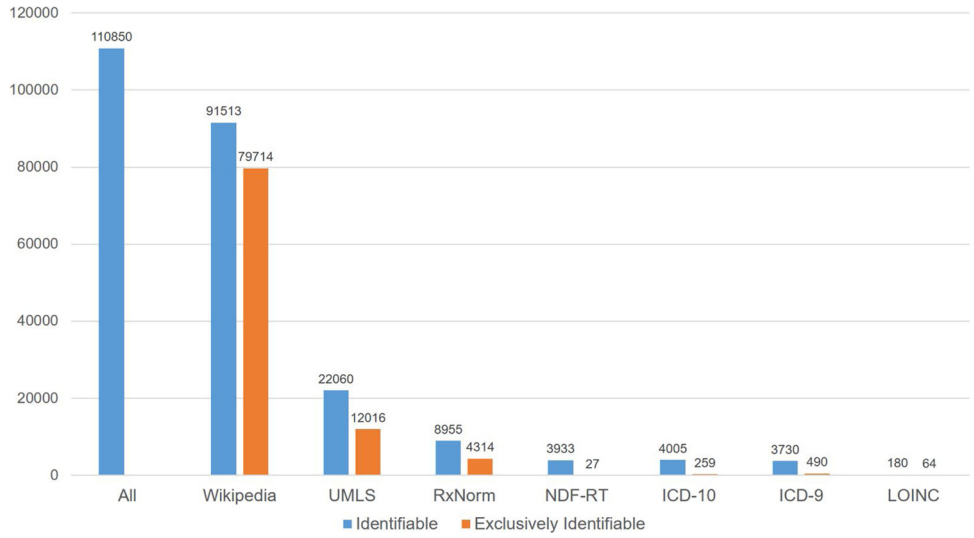

Fig. 3. The number of Wikidata items identifiable using each method.

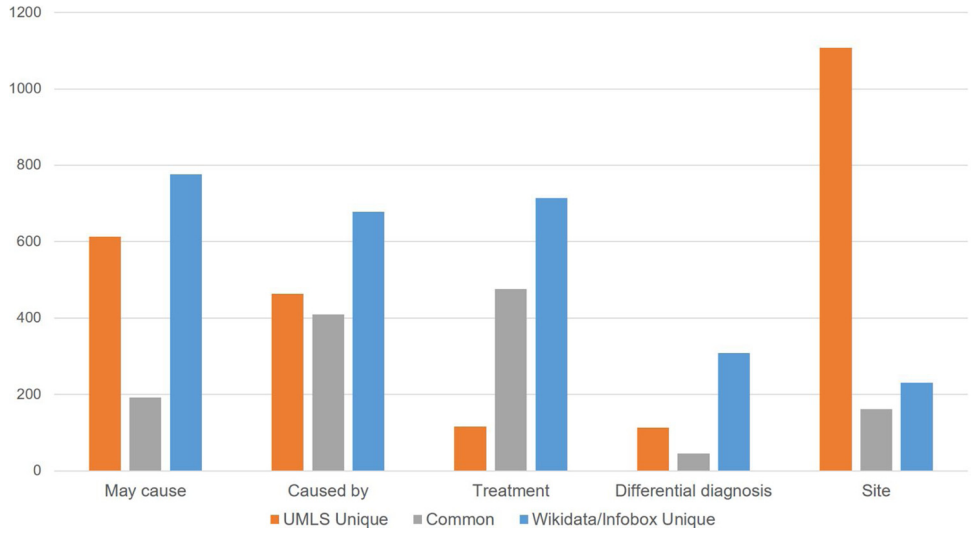

Fig. 4. Number of concepts that each relation covered that were unique in the UMLS, unique in Infobox/Wikidata, or common in both.

Table 5
Relation name mapping for May cause, Caused by, Treatment, Differential diagnosis, and Site. Words in the table show relation names used by each source; The meaning of Wikidata properties are in parentheses; 'NA': unavailable.

	UMLS relation names	Infobox relation names	Wikidata relation names
May cause	has_manifestation, has_definitional_manifestation	Symptoms, Complications	P780 (symptoms), P1542 (has effect)
Caused by	has_causative_agent, cause_of	Causes, Risk factors	P828 (has cause), P5642 (risk factor)
Treatment	may_be_treated_by	Treatment, Medication	P2176 (drug used for treatment), P924 (possible treatment)
Differential diagnosis	ddx	Differential diagnosis	NA
Site	disease_has_primary_anatomic_site, disease_has_associated_anatomic_site	NA	P689 (afflicts), P927 (anatomical location)

give us that information. Additionally, we also do not know the true ratio of non-medical articles in Wikipedia, so randomly sampling 17,000 negative samples for training is also biased. For an unbiased validation, we manually reviewed samples that were classified as medical (that is, in one of the 7 semantic groups), and results show that the proposed mechanism is far superior to the baseline text classifiers, and is the only one that has an acceptable false discovery rate (Table 4). The manual review cannot evaluate recall. If we use DISNET as a reference for diseases, then the recall is at least 95 %. Inference from this kind of positive vs. unlabeled data is an open question and active research area [44].

One of the end goals of identifying the medical subset of Wikipedia is to extract structured assertional knowledge to support the development of medical knowledge graphs. We extracted 1.3 million facts and concept names in multiple languages from the Wikidata and 667 thousand lines of properties from the Infobox, which are a wealth of information that can be used in future modeling and NLP tasks. We also wanted to know how many new diseases-related relations the extracted Infobox and Wikidata might add to what the UMLS already had. We used the recognizable UMLS DISO concepts in the identified subset that are likely diseases (see Supplementary Materials S6) as a common ground for comparison. The infobox and Wikidata are stored in a 'key:

value' format, with the key names express relations and are standardized for diseases. We investigated the relation coverage by counting what relations the UMLS, Infobox, and Wikidata included, respectively, and how many concepts each relation covered. A concept is 'covered' by a relation if its Infobox or Wikidata has the corresponding key entry. The detailed counts are given in Supplementary Materials S7. Additionally, we analyzed 5 relations that were important to clinical decision support, namely: May cause, Caused by, Treatment, Differential diagnosis, and Site. Fig. 4 compares the number of concepts that these relations covered, grouped by whether they were unique in the UMLS, unique in Infobox/Wikidata, or common in both. Table 5 gives the relation name mapping used for counting the number of concepts covered. Note that the mapped relations might not be equivalent in broadness. For example, "has causative agent" in the UMLS is a narrower relation than Caused by. From Fig. 4, one can see that Infobox and Wikidata can provide a significant supplement to the UMLS in 4 of the 5 relations. Examining closer about which diseases are covered further shows that a large proportion of those covered by Infobox/Wikidata but not by the UMLS are common diseases, such as type 2 diabetes and influenza. This could be due to that researches of some common diseases were not as heavily funded as diseases like cancer and do not have dedicated ontologies. Therefore, from the perspective of primary healthcare decision support, the value of the added relations can be more substantial than what Fig. 4 can show.

7. Conclusion

Wikipedia can provide very rich structured and unstructured information to support medical informatics. However, the subset of medical articles in Wikipedia had not been identified and the whole Wikipedia can be difficult to work with. The automatic mechanism that we developed can identify the medical articles in Wikipedia with high accuracy. In particular, the crawling classification strategy and the utilization of Wikipedia's rich structures allow it to achieve far superior performance than generic text classifiers in false discovery control. Due to the extremely low prevalence of medical articles in Wikipedia, our study is limited in the evaluation of overall recall by manually reviewed gold-standards. Our future research aims to simplify the classification process and to develop adaptive classifiers to improve the accuracy on the very short articles. To facilitate healthcare modeling and NLP, more semantic groups may be included in subsequent iterations. Additionally, automatic article quality assessment can also be added to avoid extracting knowledge from uninformative articles [45,46].

Author contribution statement

Lishan Yu is the main contributor to the programming and the result evaluation. Sheng Yu is the principal investigator of the project and is responsible for the design of the methodology. Both authors contributed significantly to drafting and critically revising this paper.

Fundings

This work was supported by the National Natural Science Foundation of China (No. 11801301), the National Key Research and Development Program of China (No. 2018YFC0910404), Beijing Natural Science Foundation (No. Z190024), and the Tsinghua University Initiative Scientific Research Program.

Summary Table

What was already known on the topic

1 Wikipedia medical articles provide comprehensive and frequently updated information that is useful for medical informatics.
2 Medical articles are extremely scarce in Wikipedia, which makes their identification with text classification difficult.

What this study added to our knowledge

1 Crawling classification is more effective than off-the-shelf text classifiers in identifying the extremely scarce medical articles from Wikipedia.
2 The extracted relations from Wikipedia infobox and Wikidata can provide rich supplement to the relations in UMLS.

Declaration of Competing Interest

None.

Acknowledgment

The authors thank the following people for their help in data collection and preliminary analyses.

Xiongyi Zhang, Department of Mathematical Sciences, Tsinghua University;

Yunan Gao, Department of Industrial Engineering, Tsinghua University;

Yucong Lin, Department of Industrial Engineering, Tsinghua University;

Zihao Fan, Department of Computer Science and Technology, Tsinghua University;

Zijie Cheng, Department of Computer Science and Technology, Tsinghua University;

Daiqi Gao, Department of Industrial Engineering, Tsinghua University;

Xinyi Zhong, Department of Foreign Languages and Literatures, Tsinghua University;

Hongyi Yuan, Department of Electrical Engineering, Tsinghua University.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104234.

References

[1] D.A. Smith, Situating Wikipedia as a health information resource in various contexts: a scoping review, PLoS One 15 (2020) e0228786, , https://doi.org/10.1371/journal.pone.0228786.
[2] Y. Cao , H. Mehta , A.E. Norcross , et al. 2020. Analysis of Wikipedia pageviews to identify popular chemicals. In: Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications XII International Society for Optics and Photonics 112560I 2020; 10.1117/12.2542835.
[3] S. Yu, K.P. Liao, S.Y. Shaw, et al., Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources, J. Am. Med. Inform. Assoc. 22 (2015) 993–1000, https://doi.org/10.1093/jamia/ocv034.
[4] A. Jagannatha, J. Chen, H. Yu, Mining and ranking biomedical synonym candidates from wikipedia, Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (2015) 142–151.
[5] G. Lagunes García, L. Prieto Santamaria, E.P. Garcia del Valle, et al., Wikipedia disease articles: an analysis of their content and evolution, 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (2019) 664–671, https://doi.org/10.1109/CBMS.2019.00136.
[6] E.P. García del Valle, G. Lagunes García, L. Prieto Santamaría, et al., Evaluating wikipedia as a source of information for disease understanding, 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS) (2018) 399–404, https://doi.org/10.1109/CBMS.2018.00076.
[7] C.B. Ahlers, M. Fiszman, D. Demner-Fushman, et al., Extracting semantic predications from medline citations for pharmacogenomics, Biocomputing 2007, WORLD SCIENTIFIC, 2006, pp. 209–220, https://doi.org/10.1142/9789812772435_0021.
[8] D.M. Lowe, N.M. O'Boyle, R.A. Sayle, Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall, Database (Oxford) (2016) 2016, https://doi.org/10.1093/database/baw039.

[9] P. Arnold, E. Rahm, Extracting semantic concept relations from wikipedia, Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), Thessaloniki, Greece: Association for Computing Machinery, 2014, pp. 1–11, , https://doi.org/10.1145/2611040.2611079.
[10] Y. Lin, C. Ma, D. Gao, et al., Long distance entity relation extraction with article structure embedding and applied to mining medical knowledge, 2019 IEEE International Conference on Healthcare Informatics (ICHI) (2019) 1–7, https://doi.org/10.1109/ICHI.2019.8904821.
[11] J.M. Abasolo, Gomez M. Melisa, An ontology-based agent for information retrieval in medicine, Proceedings of the First International Workshop on the Semantic Web (SemWeb2000 (2000) 73–82.
[12] T. Goodwin, Sm. Harabagiu, Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records, 2013 IEEE Seventh International Conference on Semantic Computing (2013) 363–370, https://doi.org/10.1109/ICSC.2013.68.
[13] A.L. Beam, B. Kompa, I. Fried, et al., Clinical Concept Embeddings Learned From Massive Sources of Multimodal Medical Data. *arXiv:180401486 [cs*, Stat] Published Online First: 4 April, (2018) (accessed 5 Oct 2018), http://arxiv.org/abs/1804.01486.
[14] F.K. Khattak, S. Jeblee, C. Pou-Prom, et al., A survey of word embeddings for clinical text, Journal of Biomedical Informatics: X 4 (2019) 100057, , https://doi.org/10.1016/j.yjbinx.2019.100057.
[15] J. Lee, W. Yoon, S. Kim, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019) btz682, https://doi.org/10.1093/bioinformatics/btz682.
[16] E. Alsentzer, J.R. Murphy, W. Boag, et al., Publicly Available Clinical BERT Embeddings, NAACL HLT 2019 (2019) 72.
[17] N.J. Temple, Wikipedia Articles on Nutrition: Are they Accurate and Complete? Curr. Nutr. Food Sci. 16 (2020) 237–240, https://doi.org/10.2174/1573401314666180327095119.
[18] H. Murray, More than 2 billion pairs of eyeballs: Why aren't you sharing medical knowledge on Wikipedia? Bmj Evid. Med. 24 (2019) 90–91, https://doi.org/10.1136/bmjebm-2018-111040.
[19] A. Azzam, D. Bresler, A. Leon, et al., Why medical schools should embrace wikipedia: final-year medical student contributions to wikipedia articles for academic credit at one school, Acad. Med. 92 (2017) 194–200, https://doi.org/10.1097/ACM.0000000000001381.
[20] N.J. Reavley, A.J. Mackinnon, A.J. Morgan, et al., Quality of Information Sources About Mental Disorders: a Comparison of Wikipedia With Centrally Controlled Web and Printed Sources. Published Online First: 1 August, (2012), https://doi.org/10.1017/S003329171100287X.
[21] G. Eysenbach, J. Powell, O. Kuss, et al., Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review, JAMA 287 (2002) 2691–2700, https://doi.org/10.1001/jama.287.20.2691.
[22] B.L. Humphreys, D.A. Lindberg, The UMLS project: making the conceptual connection between users and the information they need, Bull. Med. Libr. Assoc. 81 (1993) 170.
[23] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene Ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.
[24] P.N. Robinson, S. Köhler, S. Bauer, et al., The human phenotype ontology: a tool for annotating and analyzing human hereditary disease, Am. J. Hum. Genet. 83 (2008) 610–615, https://doi.org/10.1016/j.ajhg.2008.09.017.
[25] G. Lagunes-García, A. Rodríguez-González, L. Prieto-Santamaría, et al., DISNET: a framework for extracting phenotypic disease information from public sources, PeerJ (2020) 8, https://doi.org/10.7717/peerj.8580.
[26] C. Bizer, J. Lehmann, G. Kobilarov, et al., DBpedia - a crystallization point for the web of data. Web semantics: science, Services and Agents on the World Wide Web 7 (2009) 154–165, https://doi.org/10.1016/j.websem.2009.07.002.
[27] WikiProject Medicine. Wikipedia, (2019) (accessed 1 Jul 2020), https://en.wikipedia.org/w/index.php?title=Wikipedia:WikiProject_Medicine&oldid=911651909.
[28] J.M. Heilman, A.G. West, Wikipedia and medicine: quantifying readership, editors, and the significance of natural language, J. Med. Internet Res. 17 (2015) e62, https://doi.org/10.2196/jmir.4069.
[29] J. Devlin, M.-W. Chang, K. Lee, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] Published Online First: 10 October, (2018) (accessed 21 Oct 2018), http://arxiv.org/abs/1810.04805.
[30] Y. Liu, M. Ott, N. Goyal, et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:190711692 [cs] Published Online First: 26 July, (2019) (accessed 1 Jul 2020), http://arxiv.org/abs/1907.11692.
[31] Z. Yang, Z. Dai, Y. Yang, et al., XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:190608237 [cs] Published Online First: 2 January, (2020) (accessed 1 Jul 2020), http://arxiv.org/abs/1906.08237.
[32] C. Bunkhumpornpat, K. Sinapiromsaran, Lursinsap C. Safe-Level-SMOTE, et al., Safe-level-synthetic minority Over-sampling TEchnique for handling the class imbalanced problem, in: T. Theeramunkong, B. Kijsirikul, N. Cercone (Eds.), Advances in Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2009, pp. 475–482, , https://doi.org/10.1007/978-3-642-01307-2_43.
[33] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a New Over-sampling method in imbalanced data sets learning, in: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Springer, Berlin, Heidelberg, 2005, pp. 878–887, , https://doi.org/10.1007/11538059_91.
[34] H.M. Nguyen, E.W. Cooper, K. Kamei, Borderline over-sampling for imbalanced data classification, Int. J. Knowl. Eng. Soft Data Paradig. 3 (2011) 4–21, https://doi.org/10.1504/IJKESDP.2011.039875.
[35] M. Mintz, S. Bills, R. Snow, et al., Distant supervision for relation extraction without labeled data, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP Volume 2 - Volume 2 Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 1003–1011 (accessed 7 Mar 2016), http://dl.acm.org/citation.cfm?id=1690219.1690287.
[36] G. Angeli, J. Tibshirani, J.Y. Wu, et al., Combining distant and partial supervision for relation extraction, In Proceedings of EMNLP (2014) 1556–1567.
[37] A. Ratner, S.H. Bach, H. Ehrenberg, et al., Snorkel: rapid training data creation with weak supervision, Proceedings VLDB Endowment 11 (2017) 269–282, https://doi.org/10.14778/3157794.3157797.
[38] T. Mikolov, I. Sutskever, K. Chen, et al., Distributed representations of words and phrases and their compositionality, in: C.J.C. Burges, L. Bottou, M. Welling (Eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 3111–3119.
[39] Narrative Information Linear Extraction (NILE) Software. CELEHS. /packages/nile/ (accessed 16 Feb 2020).
[40] K. Cho, B. van Merrienboer, C. Gulcehre, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1724–1734 (accessed 8 Jun 2017), http://arxiv.org/abs/1406.1078.
[41] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification. AAAI-98 Workshop on Learning for Text Categorization, Citeseer, 1998, pp. 41–48.
[42] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765, https://doi.org/10.1016/j.eswa.2010.08.066.
[43] Y. Kim, Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1746–1751.
[44] J. Bekker, J. Davis, Learning from positive and unlabeled data: a survey, Mach. Learn. 109 (2020) 719–760, https://doi.org/10.1007/s10994-020-05877-5.
[45] E. Bassani, M. Viviani, Automatically assessing the quality of wikipedia contents, Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Association for Computing Machinery, Limassol, Cyprus, 2019, pp. 804–807, https://doi.org/10.1145/3297280.3297357.
[46] X. Li, J. Tang, T. Wang, et al., Automatically assessing wikipedia article quality by exploiting article–Editor networks, in: A. Hanbury, G. Kazai, A. Rauber (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2015, pp. 574–580, , https://doi.org/10.1007/978-3-319-16354-3_64.