Scientometrics (2021) 126:8275–8293
https://doi.org/10.1007/s11192-021-04117-4

Leveraging full-text article exploration for citation analysis

Moreno La Quatra[1] · **Luca Cagliero**[1] · **Elena Baralis**[1]

Received: 23 October 2020 / Accepted: 26 July 2021 / Published online: 18 August 2021

Abstract
Scientific articles often include in-text citations quoting from external sources. When the cited source is an article, the citation context can be analyzed by exploring the article full-text. To quickly access the key information, researchers are often interested in identifying the sections of the cited article that are most pertinent to the text surrounding the citation in the citing article. This paper first performs a data-driven analysis of the correlation between the textual content of the sections of the cited article and the text snippet where the citation is placed. The results of the correlation analysis show that the title and abstract of the cited article are likely to include content highly similar to the citing snippet. However, the subsequent sections of the paper often include cited text snippets as well. Hence, there is a need to understand the extent to which an exploration of the full-text of the cited article would be beneficial to gain insights into the citing snippet, considering also the fact that the full-text access could be restricted. To this end, we then propose a classification approach to automatically predicting whether the cited snippets in the full-text of the paper contain a significant amount of new content beyond abstract and title. The proposed approach could support researchers in leveraging full-text article exploration for citation analysis. The experiments conducted on real scientific articles show promising results: the classifier has a 90% chance to correctly distinguish between the full-text exploration and only title and abstract cases.

Keywords Citation analysis · Deep natural language processing · Citation classification

Introduction

To disseminate scientific knowledge thousands of papers have been published every day in national and international journals, conference proceedings, and books. Scientists commonly explore the literature related to their research area prior to writing a scientific article.

✉ Moreno La Quatra
moreno.laquatra@polito.it

Luca Cagliero
luca.cagliero@polito.it

Elena Baralis
elena.baralis@polito.it

[1] Politecnico di Torino Corso Duca degli Abruzzi, 24, 10129 Turin, Italy

Even though, in most cases, they know in advance the publication venues where relevant previous works had been published, performing an exhaustive literature review could be extremely time-consuming.

To reduce the time spent perusing the related literature, a shortcut is to browse the citation network. For example, starting from few, authoritative papers, readers could explore the cited publications in order to deepen the study of a particular topic. Exploring the cited paper is typically useful when the citing context of a citation (i.e., the snippet of text in the citing paper surrounding the citation) is not self-explanatory or when there is a need to gain additional knowledge on that particular aspect. However, reading the full-text of the cited articles could be still time-consuming, Therefore, it is worth exploring the textual correlation between the separate sections of the cited paper and the local context where the citation is placed (He et al., 2010). The aim is to identify the sections that are worth reading since their content is likely to be correlated to that of the citing context.

Given a large corpus of annotated academic papers [i.e., the ScisummNet dataset (Yasunaga et al., 2019)], we first perform a data-driven correlation analysis between the citing contexts (namely the citances) and the content of the cited articles. Specifically, we explore the relationship between the text of the citance and the content of the separate sections of the cited paper. The goal is to identify the sections that are most likely to include pertinent content. To this aim, we train explainable classification models trained on citance-cited paper section pairs in order to capture the most significant underlying text correlations. To this end, we describe citance-cited section relationships according to various syntactic and semantic features, including those trained using Deep NLP models (Mikolov et al., 2013; Pagliardini et al., 2018; Devlin et al., 2019). The experiments carried out on 1000 papers and 16,981 citations have shown that

- the content of title and abstract of the cited paper is highly similar to that of the citing context. Notice that for most of the academic publications title and abstract are accessible without paying any editorial fee, whereas the access to the subsequent sections can be restricted.
- for a subset of the analyzed citations it is worth exploring also method, experiments, or conclusion sections because their content has shown to provide relevant information about the citing context.

A citance often points to several text snippets in the cited paper. When the cited text is in both the title and abstract and in the subsequent sections of the paper it is unclear the extent to which a full-text exploration of the cited paper provides additional knowledge on the citance beyond reading only the title and abstract. Despite full-text article exploration definitely provides additional knowledge, there is a need to understand the extent to which an exploration of the full-text of the cited article would be beneficial to gain insights into the citing snippet, considering also the fact that the full-text access could be restricted. Specifically, given a citance and the preliminary (freely accessible) sections of a cited paper, we aim at understanding whether a machine learning-based approach could support researchers in automatically identifying the citances that would require further explorations of the cited article beyond simply reading the title and abstract. In other words, the main goal of the present study is to use classification techniques to leverage full-text article exploration for citation analysis.

To address the aforesaid issue, we design a classification approach trained on citance-cited paper pairs extracted from ScisummNet data (Yasunaga et al., 2019, 2017). The training dataset includes a selection of features describing the similarity between the content

Table 1 Key statistics on the ScisummNet dataset

Statistic	Value
Num. of papers	998
Num. of citations	16981
Min. num. of citations per paper	3
Max. num. of citations per paper	20
Avg. num. of citations per paper	17.01
Min. abstract length [num. of words]	21
Max. abstract length [num. of words]	964
Avg. abstract length [num. of words]	112.15
Min. full-text length [num. of words]	316
Max. full-text length [num. of words]	22821
Avg. full-text length [num. of words]	4438.94

of the preliminary sections of the cited paper and that of the citance. The key idea is to analyze the textual correlations among the title and abstract of the cited paper and citances in order to automatically recommend a full-text exploration of the cited paper.

The experimental results achieved on the ScisummNet dataset show promising performance: the machine learning models have a 90% chance to correctly distinguish between full-text and only title-abstract cases.

The rest of the paper is organized as follows. Section 2 describes the ScisummNet data collection. Section 3 details the research questions addressed by the current work. Section 4 overviews the related literature. Section 5 describes the features used to model the similarity between citance and cited text. Section 6 presents the correlation analysis and summarizes the key results, whereas in Sect. 7 the classification approach is described and an empirical validation of the proposed solution is reported as well. Finally, Sect. 8 draws conclusions and discusses the future developments of this work.

The ScisummNet dataset

CL-SciSumm (Chandrasekaran et al., 2019) is an yearly research challenge focused on the analysis of scientific paper full-text and scientometric data. More specifically, it addresses the analysis of a set of topics, where each topic consists of a reference paper (RP) and a set of citing papers (CPs), all containing citations to the RP.

Since 2019 the CL-SciSumm Shared Tasks have been carried out on the ScisummNet data collection (Yasunaga et al., 2019, 2017). ScisummNet is a large-scale, human-annotated dataset consisting of about 1000 papers in the ACL anthology network with their citation networks (e.g. citation sentences, citation counts) and their comprehensive, manual summaries.[1] ScisummNet was manually annotated by the contest organizers to foster the study and development of innovative research in citation-aware scientific paper summarization.

[1] ScisummNet project link: https://cs.stanford.edu/~myasu/projects/scisumm_net/.

Table 2 Notation used throughout the paper

Symbol	Description
P	Scientific paper collection
p_i	Citing paper
p_j	Cited paper
cit	Citance, i.e., a sentence in p_i including the citation.
cts	Cited text span, i.e., a snippet of text referenced by cit
$ns(p_i)$	number of sections in paper p_i
$S^k(p_j)$	The k-th section of cited paper p_j
α	Similarity scores' weight
$t + a$	Title and abstract sections

A selection of statistics on the analyzed dataset is reported in Table 1. Notably, the number of citations per paper, the abstract and full-text sizes are quite variable thus highlighting the complexity of the addressed task.[2]

Let P be the set of scientific papers enriched with citation information. For each citation ScisummNet stores (i) the full-text of the citing paper in $p_i \in P$ (including title and abstract, but excluding funding and acknowledgement information), (ii) the snippet of text (typically one sentence) in p_i including the citation, namely the *citance* (*cit*), (iii) the full-text of the cited paper $p_j \in P$, (iv) the snippets of text (typically few consecutive sentences) in p_j to which the citance refers to, namely the *cited text snippets* (*cts*). Notice that each citance can be mapped to one or more cited text snippets (typically from one to five).

Since 2016 the CL-SciSumm research challenges consist of the following Shared Tasks:

- *Task 1A:* For each citance, identify the cited text spans in the RP that most accurately reflect the citance.
- *Task 1B:* For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets.
- *Task 2:* Generate a structured, fixed-length summary of the RP from the cited text spans of the RP.

The present study is grounded on scientometric data available in the ScisummNet data set. It addresses a new research challenge partly related to Tasks 1A and 1B, but conceptually different from those previously addressed by CL-SciSumm, namely the use of machine learning to leverage full-text article exploration for citation analysis. A thorough description of the addressed problem is given in Sect. 3.

[2] Few tens of scientific papers in the ScisummNet collection were disregarded in the word counts due to clear inconsistencies in raw text parsing.

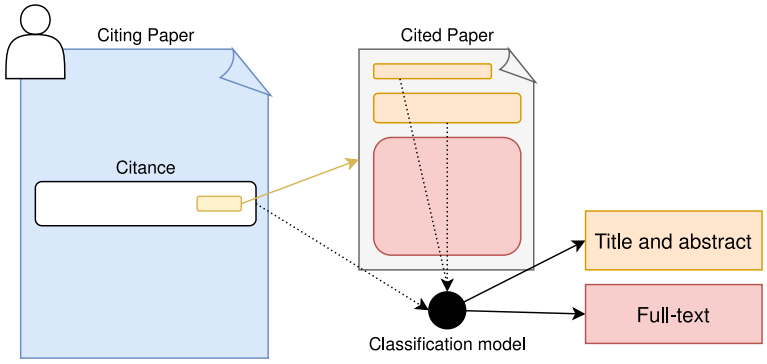


Fig. 1 Use case sketch

Problem statement and practical use case

We hereafter formalize the problem addressed by the present paper on the ScisummNet collection (Yasunaga et al., 2019). Notice that the same problem can be trivially extended to similar data collections. For our convenience, hereafter we will rely on the notation summarized in Table 2.

We analyze a collection of papers, where the text snippet within a citance *cit* in p_i contains a citation to a specific cited paper $p_j \in P$. For the sake of simplicity, citations within the same text span are deemed as distinct citances.

Based on the paper structure, the content of the cited paper can be partitioned into sections. Let $S^k(p_j)$ be the *k-th* section of p_j. Common examples of sections are *title*, *abstract*, *introduction*, *methodology*, *experiments*, and *conclusions*. *Title* and *abstract* can be approximately classified as *open-access*, since most editors give free access to their content. Conversely, all the remaining sections will be hereafter denoted by *restricted* since their content (depending upon the editor's policies) could be not freely available.[3]

The goal of this paper is to address the following research questions.

RQ1 Explore the correlation between the textual content of the citance $cit \in p_i$ and that of each section $S^k(p_j)$ $[1 \leq k \leq ns(p_i)]$.

RQ2 Classify citances as (i) *open-access*, if the preliminary, open access sections of the cited paper (i.e., title and abstract) provide most of the related information, or (ii) *restricted*, if relevant information is provided by the subsequent sections.

The former task (RQ1) entails studying the syntactic and semantic relationships between citances and cited sections. It provides insights into the pertinence between the citing text span and different portions of the cited text. The latter task (RQ2) focuses on automatically classifying a citance as either mainly described title and abstract (i.e., further explorations of the full-text of the cited paper could be not necessary) or described by other (potentially restricted) content not present in the title and abstract.

[3] For the sake of simplicity, we will ignore the fact that the content of some papers is fully available for free since such information is not available in the ScisummNet dataset.

The proposed methodology has a typical use case, which is sketched in Fig. 1 and briefly described below.

Use case description A researcher is reading a paper, which contains a citation. She/he would like to deepen his knowledge on the main topic covered by the citance (i.e., the text around the citation). She/he could access the editorial version of the cited paper and read its introductory parts (i.e., title and abstract). We aim at proposing a classification-based system that supports the researcher in deciding whether to further explore the article full-text or not. The classification method automatically predicts whether the content of the remaining sections of the cited paper is likely to include a relevant amount of additional knowledge compared to the title and abstract.

Related works

We discuss the position of the present work compared to the previous contributions to the existing CL-SciSumm tasks (see Sect. 4.1), the previous studies on citation-based paper indexing (see Sect. 4.2), and the works related to citation classification and recommendation (see Sect. 4.3).

Contributions to CL-SciSumm

The CL-SciSumm research challenge has fostered cutting-edge research on scientometric data. To identify the cited text spans (i.e., task 1a), most previous works focused on modeling the relation between citances and candidate cited sentences. For example, the best performing approach presented in the latest edition (2019) exploited pre-trained sentence embeddings to capture text similarity using Deep NLP techniques (Zerva et al., 2020). Alternative approaches have exploited CNN and LSTM Neural Network architectures (e.g., AbuRa'ed et al., 2018; Moraes et al., 2018; Nomoto, 2018), Latent Dirichlet Allocation (e.g., Li et al., 2018), and similarity-based models relying on a mix of citation-dependent and citation-independent features (e.g., La Quatra et al., 2019; Yeh et al., 2017).

According to a pilot study of the biomedical summarization track of the text analysis conference 2014,[4] most citations refer to one or more specific aspects of the cited paper (usually *Aim*, *Method*, *Result/Data*, or *Conclusion*) (Ronzano and Saggion, 2016). Hence, an appealing parallel research direction is the automatic identification of the citation discourse facet (i.e., task 1b). Most related works have adopted binary classifiers combining various latent features trained using Deep NLP techniques (e.g., Davoodi et al., 2018; Ma et al., 2018; Wang et al., 2018). Other approaches have considered simpler word- or sentence-based relevance scores (e.g., Baruah et al., 2018; Li et al., 2018). Recently, facet annotations have also been exploited to produce discourse facet summaries of scientific papers (La Quatra et al., 2020).

A more detailed overview of the latest CL-SciSumm contributions is given in Chandrasekaran et al. (2019). This work extends the original CL-SciSumm tasks by exploring the correlation between the content of the text snippet where the citation is placed and the sections of the cited article content. Unlike task 1.b, we analyze text at the section-level (instead of at the sentence-level) and the goal is not faceted classification. To the best of

[4] https://tac.nist.gov/2014/BiomedSumm/index.html.

our knowledge, exploring section relevance to a particular citance cannot be addressed by any existing approaches addressing the CL-SciSumm tasks.

Citation-based paper indexing

Exploring the citation network is crucial for gaining knowledge on specific scientific topics. Under this umbrella, a relevant research problem entails effectively retrieving the papers that are most pertinent to a given citation. For example, the authors in Ritchie et al. (2006) studied how exploring the textual context of the citations in scientific papers could improve the indexing of the cited papers. They studied the effect of combining the existing index terms of a paper with additional expressions from citing articles. Related studies on using terms around citations to index the cited paper have also been presented in Ritchie et al. (2008, 2008). All the aforesaid studies confirmed the importance of correctly selecting and deeply analyzing the text around citations (i.e., the citances) to gain insights into scientometric data (Khalid et al., 2017; Ritchie et al., 2008). However, to our best knowledge, they do not differentiate between the citances whose content is strongly correlated to specific sections of the cited paper, which is instead the main goal of the present work.

Citation classification and recommendation

To enrich scientific papers with scientometric data, the corresponding citations can be automatically annotated. For example, in Jha et al. (2017) and Yousif et al. (2019) the authors focused on predicting the sentiment (polarity) of a citation as well as its main purpose. The relevance of a citation to a specific context has been investigated in Hernandez-Alvarez et al. (2017), whereas in Cohan et al. (2019) and Jurgens et al. (2018) the authors focused on inferring the citation intent.

Automated citation classification is typically the first step of a personalized citation recommender, whose goal is to support researchers in finding a relevant manuscript on the web according to their actual needs. Citation recommendation entails identifying the citations that are most pertinent to a given academic paper (Jeong et al., 2019). When the citation must be pertinent to a particular snippet of text where the citation should be placed, the problem is commonly denoted by *context-aware* citation recommendation (Jeong et al., 2020; Cohan et al., 2020). A recent survey on citation recommendation is given in Ali et al. (2020). The problem addressed by this paper also falls into the citation classification, but the aim of the present work is substantially different from all the aforesaid ones.

Feature engineering

To explore the relationship between a citance and different sections of the cited papers in the ScisummNet collection (Yasunaga et al., 2019), we first map the original section headers to a predefined subset of categories, namely *Title*, *Abstract*, *Introduction*, *Related works*, *Method*, *Experiments* and *Conclusions and future works* using English-based regular expressions. Then, we extract a subset of features describing the syntactic and semantic relationships between the content of each section of the cited paper and the text snippet in the citance.

A more detailed description of the analyzed features is given below, whereas Table 3 summarizes the naming convention used throughout the paper.

Table 3 Name conventions used throughout the paper

Abbreviation	Description
S2V	Cosine similarity between Sent2Vec embeddings
W2V	Cosine similarity between Word2Vec embeddings
BERT	Cosine similarity between BERT embeddings
R1	Rouge-1 score
R2	Rouge-2 score
RL	Rouge-L score
R*	Rouge-based methods (R1, R2, RL)
V*	Embedding based methods (S2V, W2V)
F	Full set of features (R+V)
T	Title section
A	Abstract section
I	Introduction section
R	Related works section
M	Method section
E	Experiments section
C	Conclusions and future works section

Features derived from vector representations of text. In recent years, vector representations of text have become extremely popular in text mining and analytics. They have found application, for instance, in speech recognition, automatic text translation, and analogy detection (Zhang et al., 2018). The use of high-dimensional text encodings allows us to capture linguistic regularities and semantic relationships between citances and cited sections.

In the subsequent analyses we consider the following embedding models: (1) a popular word-level embedding model, i.e., Word2Vec (Mikolov et al., 2013). (2) a contextualized sentence-level embedding model, namely Sent2Vec (Pagliardini et al., 2018). (3) a transformer-based model, namely BERT (Devlin et al., 2019), which leverages the attention mechanism to attend relevant information at the sentence level. Word-level embeddings, such as (1), suffer from the lack of word contextualization. To derive the sentence encoding, the corresponding vectors are averaged (Pilehvar and Camacho-Collados, 2020). Contextualized embeddings, such as (2) and (3), provide a dynamic word contextualization. Among them, BERT-based transformers have recently outperformed all the other models in several NLP tasks.

To capture word- and sentence-level text similarities, we compute the similarity between the sentence-level representations of the cited and citing text separately for Word2Vec (Mikolov et al., 2013), Sent2Vec (Pagliardini et al., 2018), and BERT (Devlin et al., 2019).[5]

Features derived from co-occurrence-based models. We compare also the content of each section with that of the citing context using the established Recall-Oriented Understudy for Gisting Evaluation (Rouge) toolkit (Lin, 2004). Rouge measures the unit overlap between two different text snippets. It quantifies the syntactic similarity between two

[5] We used the Sentence-BERT model (Reimers and Gurevych, 2019), which is specifically tuned for the semantic similarity task.

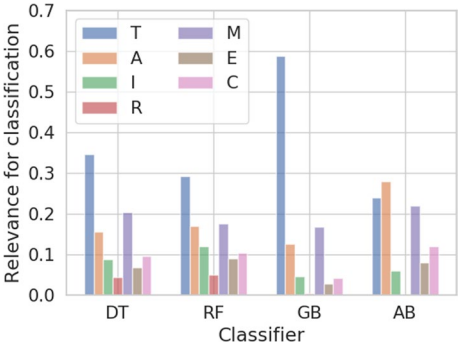


Fig. 2 Section-level feature relevance comparison. ScisummNet dataset

portions of text by looking for the n-grams in common between the two. To our purposes, hereafter we will rely on the Rouge recall measure, which is given by

$$R_n = \frac{C_{match}(n_{gram})}{C(n_{gram})}$$

where $C(n_{gram})$ is the n-grams' count in the reference text whereas $C_{match}(n_{gram})$ is the number of matching n-grams in the tested text snippets. The idea behind it is to verify the coherence, in terms of percentage of common n-grams, between the tested and reference text snippets. In this work, we target the Rouge recall values expressed in terms of unigrams (Rouge-1), bigrams (Rouge-2), and longest common subsequence (Rouge-L), respectively.

Correlation analysis

To answer the Research Question 1, we carry out a data-driven text correlation analysis between the citances and the sections in the cited paper. To this purpose, we build a classification model aimed at predicting whether an arbitrary citance points to a specific paper. Classifier decisions are taken based on a set of features describing the textual similarity between the citance and each section of the candidate paper. The idea behind it is to *explain* classifier decisions in order to understand which sections of the cited papers are most discriminating in predicting the correct mapping between citances and cited papers.

We model the features extracted from the raw textual content as a labeled relational dataset (Zaki et al., 2020). Specifically, each record of the dataset is identified by a distinct pair $\langle\, p_j, cit \,\rangle$, where p_j is an arbitrary paper and cit is a citance. For each feature described in Sect. 5, we generate a separate dataset attribute separately for each section in p_j. Each record has a label (l) that indicates whether cit cites p_j or not. Specifically, if the citation included in cit points to the paper p_j then the label is positive (negative otherwise). Since each citance references only one paper, to avoid introducing a bias due to class imbalance we perform under-sampling of the negative class.

We train the following classification models: Decision Trees, Random Forest, Gradient Boosting, and AdaBoost (Zaki et al., 2020). They are deemed as particularly suitable for explaining the relationship between citances and cited paper sections because they provide an estimate of the feature relevance score in classification thus indicating which section-level descriptors are most discriminating.

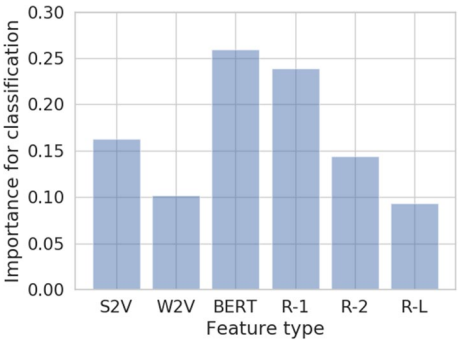


Fig. 3 Feature relevance analysis. ScisummNet dataset

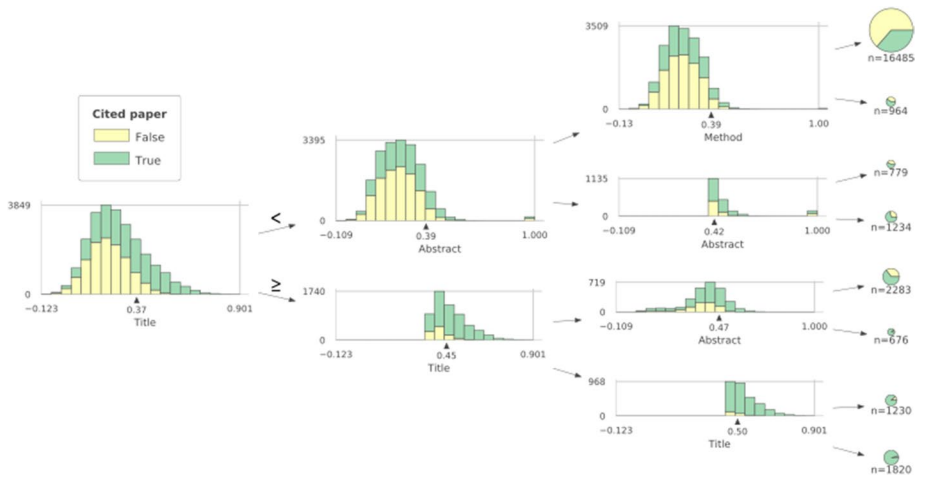


Fig. 4 Example of Decision tree model. ScisummNet dataset

Figure 2 summarizes the feature relevance scores, averaged over the sections of the cited papers, separately for each algorithm. Independently of the considered model, *title*, *abstract*, and *method* appeared to be the most discriminating sections.

To further investigate the importance of each category of features in the classification phase, we also compute the aggregated feature importance scores separately per feature category. Figure 3 reports the importance level of each feature type, according to a representative classifier (Decision Tree), separately for each section. The results shows that BERT-based features are the most discriminating ones.

To gain insights into the characteristics of the trained models, Fig. 4 shows an example of decision model trained on the relational dataset.[6] Decision trees are popular classification models relying on tree-based structures. Due to their inherent simplicity, decision trees can be manually explored to understand the rationale behind class label assignments. The root and intermediate nodes of the decision tree are the non-predictive data features, while

[6] For the sake of readability, we considered only the most discriminating feature type, i.e., the BERT features.

each leaf node is associated with a class label (Zaki et al., 2020). At each intermediate node, a test condition is applied to the test record on the value of the corresponding non-predictive feature. The test outcome is assigned based on one or more cutoff thresholds. For instance, the most discriminating feature (depicted in the left hand-side of Fig. 4) is the BERT-based similarity computed between the title of the cited paper and the citing context. Once a test record has to be classified, the test on this particular feature value is applied first to decide which branch of the tree must be visited further. Hence, to predict whether a section is relevant to a given citance, the decision tree model recommends us to evaluate first the similarity with the title of the cited paper in order to take a decision.

Based on the outcomes of the test performed at the first stage, different branches of the tree can be visited further. For example, according to the tree in Fig. 4, the similarity with the title and abstract are considered at the second stage. The histogram depicted within each node of the tree shows the distribution of the BERT similarity values, the cutoff threshold used for the test (0.37 for the root node), and the percentages of training records belonging to the positive and negative class labels, respectively, that fall in each interval. The top-down tree visit stops when all the remaining records (or the majority of them) have the same class. In the reported example, the stop of the tree visit is forced at the third stage for the sake of readability.

Recommending full-text article exploration for citation analysis

Given the additional knowledge provided by the full-text article content for citation analysis, we are interested in finding the cases when the contribution of the additional knowledge is less significant. Specifically, the goal of the Research Question 2 is to automatically identify the cases in which the full-text of the cited paper is worth considering beyond title and abstract content. To this aim, we propose a classification method and apply it to the ScisummNet collection (Yasunaga et al., 2019). Notice that, thanks to the generality of the proposed approach, it can be trivially applied to similar data collections as well.

The classifier takes as input (i) a citance *cit* and (ii) the content of the *title* and *abstract* sections (*t+a*, in short) of the cited paper p_j pointed by *cit*. The classification model is trained on a relational dataset describing the relationship between the textual content of citance and that of *t+a*. Each record of the training dataset is identified by a distinct pair $\langle p_j, cit \rangle$. The features used to describe the similarity relationship are summarized in Sect. 5.

Let cts_{t+a} be the subset of cited text spans in *cts* that occur in *t+a*. Let cts_{full} be the remaining cited text spans in p_j, i.e., $cts = cts_{t+a} \cup cts_{full}$. Each record has a label (l) indicating the overlap between cts_{t+a} and cts_{full}. Specifically, in the training set we label a record $\langle p_j, cit \rangle$ as *positive* if

- the cited text snippets *cts* in p_j are all located in *t+a*, i.e., $cts = cts_{t+a}$ or
- at least one of the cited text snippets in p_j is located in *t+a* (i.e., $cts_{t+a} \neq \emptyset$) and the remaining cited text snippets in *cts* (cts_{full}) are *highly similar* to *t+a*.

Otherwise, the record is labeled as *negative*.

The idea behind it is to first verify the presence of cited text spans in the introductory sections of p_j (*t+a*) and then compare the content of *t+a* with that of the cited text spans in the remaining parts of the cited paper. If the classifier predicts *positive* then the exploration

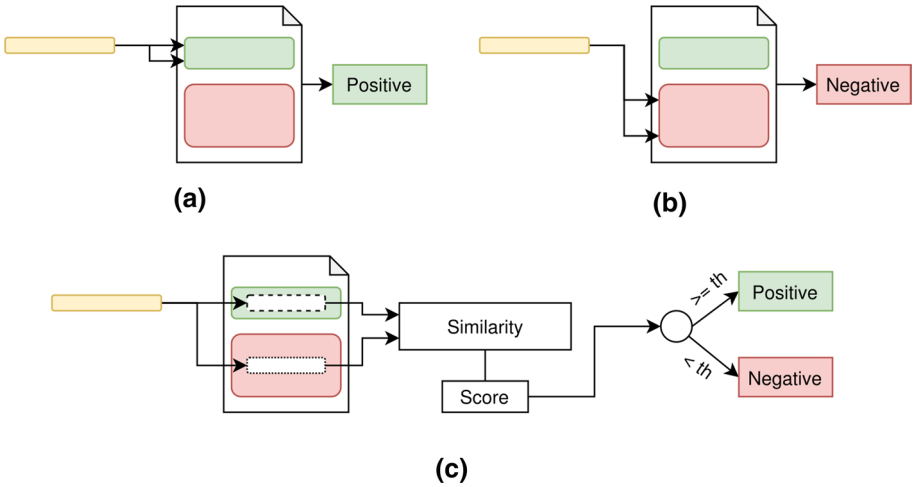


Fig. 5 Sketch of the labelling procedure

of the full-text of the cited paper is not recommended since the cited text spans located in the subsequent sections does not provide sufficiently new information compared to that already available in *t+a*. Conversely, if the classifier predicts *negative* then a significant portion of the cited content is not present in *t+a*. Therefore, the full-text of the cited paper is worth exploring.

We quantify the level of similarity between cts_{full} and *t+a* according to the following complementary aspects: (i) syntactic similarity, which relies on co-occurrence-based models and (ii) semantic similarity, which is expressed by the similarity between the vector representations of text in the latent space (Mikolov et al., 2013). More specifically, hereafter we will consider the following similarity measures:

- *Syntactic similarity (synsim)*: n-gram co-occurrences in cts_{full} and *t+a*, computed in terms of Rouge-2 Recall measure.
- *Semantic similarity (semsim)*: similarity in the embedding latent space between cts_{full} and *t+a*, computed by the cosine similarity between the corresponding BERT sentence vectors.
- *Mixed similarity*: a mix of the above similarity scores:

$$mixedsim = \alpha \cdot synsim + (1 - \alpha) \cdot semsym$$

where α ($0 \leq \alpha \leq 1$) expresses the importance of syntactic and semantic similarity scores.

cts_{full} and *t+a* are deemed as highly similar if their similarity score is above 70%. A sketch of the data labeling procedure is depicted in Fig. 5.

Quantitative empirical evaluation

We empirically evaluated the ability of the classification algorithms to correctly predict the target class (i.e., *positive* or *negative*) of an arbitrary pair of citance and cited paper

Table 4 Classifiers' performance. ScisummNet dataset. $\alpha = 0.5$

Classifier	AUC	Accuracy	Negative class		
			Prec	Recall	F1-Score
AdaBoost	0.90	0.90	0.92	0.97	0.94
GNB	0.86	0.87	0.93	0.92	0.93
GB	0.91	0.91	0.94	0.96	0.95
MLP	0.91	0.90	0.93	0.96	0.94
DT	0.71	0.87	0.93	0.93	0.93
RF	0.90	0.91	0.93	0.97	0.95

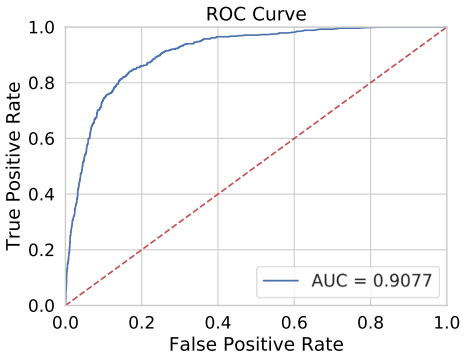
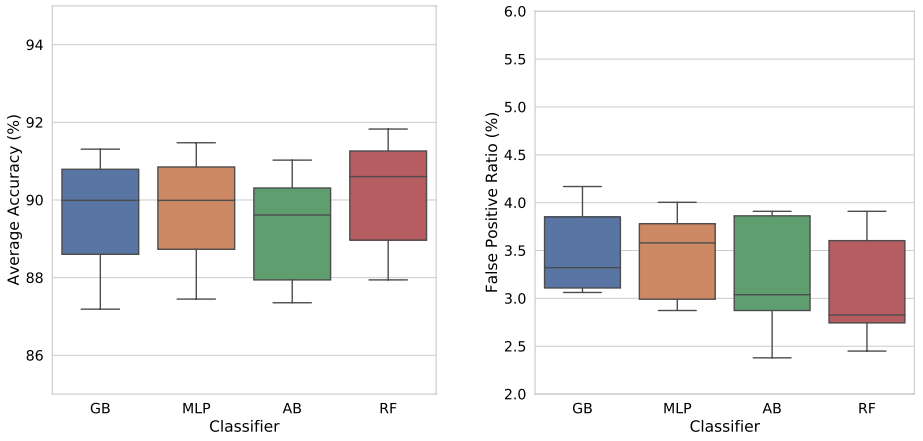


Fig. 6 Receiver operating characteristic. ScisummNet dataset. Gradient boosting classifier

$\langle p_j, cit \rangle$. Notice that, for our purposes, the two class labels are not equally important: a record labeled as *positive* corresponds to a citation that does not require a full-text cited paper exploration. Hence, the corresponding human effort is quite limited. Conversely, a record labeled as *negative* entails exploring the full-text of the cited paper thus the effort devoted to text perusal is more significant. Therefore, we deem a false negative classification error as more critical than a false positive one. To evaluate classifier performance, we conducted an empirical campaign using a variety of different classification algorithms, i.e., AdaBoost, decision trees (DT), gradient boosting (GB), multilayer perceptron (MLP), and Gaussian Naive Bayes (GNB). To train the classifiers we exploited the implementations available in the Scikit Learn Python library (Pedregosa et al., 2011). All the experiments were run on a machine equipped with Intel® Xeon® X5650, 32 GB of RAM and running Ubuntu 18.04.1 LTS.

The results are summarized in Table 4 and in Fig. 6. Specifically, in Table 4 we report the results according to the following evaluation measures:

- The *Area under curve (AUC) of the receiver operating characteristic (ROC)*, which evaluates the ability of the classifier to correctly discriminate between positive and negative labels. Specifically, it indicates the probability that the classifier outcome applied to a positive record (i.e., a citation for which a full-text exploration is not necessary) is higher than those applied to a negative record (i.e., a citation requiring a full-text cited paper exploration).
- *Accuracy*: it indicates the fraction of correctly classified records.
- *Precision of class negative*: it indicates the fraction of correctly classified negative records among all the records labeled as negative.



(a) Accuracy comparison. **(b)** False positive ratio comparison.

Fig. 7 Effects of varying the α parameter

- *Recall of class negative*: it indicates the fraction of negative records that have been retrieved over the total number of negative records.
- The *F1-score of class negative*: it is the harmonic mean of precision and recall of class *negative*.

To investigate the ability of the classifier to correctly handle the most critical situations (i.e., when a classifier error would entail a relevant human effort), we focus the per-class evaluation on precision, recall, and F1-score of the negative class.

For each classifier we performed a grid search and in Table 4 we reported the values achieved separately for each evaluation metric. Figure 6 plots the ROC curve for the best performing classifier (Gradient Boosting). All the tested classifiers show a fairly good trade-off between sensitivity (true positive rate) and specificity (inverse of the false positive rate). The best performing classifiers achieved 91% AUC, meaning that there is 91% chance that the classification model will be able to correctly distinguish between positive and negative cases. The achieved F1-score values indicate that the ability of the classifiers to correctly predict the negative cases is very high (around 94%). The choice of the classification algorithm and of the configuration setting slightly affects classification performance.

Impact of the similarity score

We explored also the effect of the type of similarity score used to compare citances and cited text snippets. Specifically, we varied the value of the α parameter to weigh differently the importance of syntactic (Rouge-based) and semantic (BERT-based) scores.

The box-plots in Fig. 7 respectively show the variations of average percentage accuracy and false positive ratio (FPR) of different classifiers by testing several α values in the range [0,1]. The results show limited performance variability (e.g., within 0.5% FPR

Table 5 Examples of positive/negative predictions. ScisummNet dataset

Label	Type	Text
Positive	Citance (citing paper)	Other attempts to address *efficiency* include the fast *Transformation based learning* (TBL) Toolkit (Ngai and Florian,2001) which *dramatically speeds up training TBL* systems, and the translation of TBL rules into finite state machines for very fast tagging
	Title (cited paper)	Transformation based learning in the fast lane
	Abstract (cited paper)	*Transformation-based learning* has been successfully employed to solve many natural language processing problems. It achieves state-of-the-art performance on many natural language processing tasks and does not overtrain easily. However, it does have a serious drawback: the training time is often intolerably long, especially on the large corpora which are often used in NLP. In this paper, we present a novel and realistic method for *speeding up the training time* of a transformation-based learner without sacrificing performance. The paper compares and contrasts the training time needed and performance achieved by our modified learner with two other systems: a standard transformation-based learner, and the ICA system. The results of these experiments show that our system is able to achieve a *significant improvement in training time* while still achieving the same performance as a standard transformation-based learner. This is a valuable contribution to systems and algorithms which utilize transformation-based learning at any part of the execution.
Negative	Citance (citing paper)	Adaptor Grammars are formally defined in Johnson et al (2007b), which should be consulted for technical details.
	Title (cited paper)	Bayesian Inference for PCFGs via Markov chain Monte Carlo
	Abstract (cited paper)	This paper presents two Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of probabilistic con- text free grammars (PCFGs) from terminal strings, providing an alternative to maximum-likelihood estimation using the Inside-Outside algorithm. We illustrate these methods by estimating a sparse grammar describing the morphology of the Bantu language Sesotho, demonstrating that with suitable priors Bayesian techniques can infer linguistic structure in situations where maximum likelihood methods such as the Inside-Outside algorithm only produce a trivial grammar

variations for all the tested classifiers). In the recommended α setting, the two contributions are equalized ($\alpha = 0.5$).

Qualitative empirical evaluation

We explored the outcomes of the classification process in order to gain insights into the automated labeling process. Table 5 reports two citation examples, respectively belonging to the positive and negative cases. In the positive case, the citance refers to a particular technique, i.e., transformation based learning (TBL), and to the efficiency problem addressed by the paper. The title explicitly mentions the name of the considered technique, whereas the abstract details the goal of the paper, i.e., speed up the TBL training process. Overall, the title and abstract include highly similar content compared to the citance.

In the negative sample the citance includes a reference to a paper to be consulted for technical details about Adaptor Grammars. However, the title and abstract of the cited paper do not mention Adaptor Grammar at all. Regarding the citation, the content of the title and abstract seems to be not self-explanatory. Thus, researchers who are interesting in getting insights into that particular topic probably need to explore the remaining sections of the cited paper.

To allow comparative analyses and foster further research on the same dataset, we made the full set of classifier outcomes available.[7]

Conclusions and future works

The paper presents a classification-based approach to analyzing the textual correlation between the section-level content of cited papers and the context in which citations are placed in the citing papers (i.e., the text around the citation). By exploring explainable classification models, we got interesting insights into the correlations hidden in the analyzed textual data.

Full-text article exploration definitely provides additional knowledge. However, it is unclear the extent to which an exploration of the paper sections beyond title and abstract is beneficial to gain insights into the citing snippet. The empirical results show that a classification model is able to accurately discriminate between the cases showing a clear benefit and not.

The achieved results leave room for various extensions. Firstly, we plan to extend the proposed methodology to open scientometric data sets where links between citances and cited text spans are not explicit [e.g., Saier and Färber (2020)]. Secondly, we plan to explore the applicability of DNN architectures to accomplish the same task. Finally, we would like to design a citation recommender that takes into account the text relationships at the section level as well as the characteristics of the paper content in terms of availability (i.e., open-access vs. restricted). The developed citation recommender system can be integrated into an existing reviewer assignment tool (Cagliero et al., 2021).

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

[7] https://github.com/MorenoLaQuatra/scim-fte

References

AbuRa'ed, A., Bravo, À., Chiruzzo, L., & Saggion, H. (2018). Lastus/taln+inco @ cl-scisumm 2018: Using regression and convolutions for cross-document semantic linking and summarization of scholarly literature. In: P. Mayr, M.K. Chandrasekaran, K. Jaidka (Eds.) Proceedings of the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2018) co-located with the 41st international ACM SIGIR conference on research and development in information retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018, *CEUR Workshop Proceedings* (Vol. 2132, pp. 150–163). http://ceur-ws.org/. http://ceur-ws.org/Vol-2132/paper15.pdf.

Ali, Z., Kefalas, P., Muhammad, K., Ali, B., & Imran, M. (2020). Deep learning in citation recommendation models survey. *Expert Systems with Applications 162*, 113790. https://doi.org/10.1016/j.eswa.2020.113790. http://www.sciencedirect.com/science/article/pii/S0957417420306126.

Baruah, G., & Kolla, M. (2018). Klick labs at cl-scisumm 2018. In P. Mayr, M.K. Chandrasekaran, & K. Jaidka (Eds.) Proceedings of the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2018) co-located with the 41st international ACM SIGIR conference on research and development in information retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018, *CEUR Workshop Proceedings* (Vol. 2132, Pp. 134–141). http://ceur-ws.org/. http://ceur-ws.org/Vol-2132/paper13.pdf.

Cagliero, L., Garza, P., Pasini, A., & Baralis, E. (2021). Additional reviewer assignment by means of weighted association rules. *IEEE Transactions on Emerging Topics in Computing, 9*(1), 329–341. https://doi.org/10.1109/TETC.2018.2861214.

Chandrasekaran, M.K., Yasunaga, M., Radev, D.R., Freitag, D., & Kan, M. (2019). Overview and results: Cl-scisumm shared task 2019. In Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2019) co-located with the 42nd international ACM SIGIR conference on research and development in information retrieval (SIGIR 2019), Paris, France, July 25, 2019 (pp. 153–166). http://ceur-ws.org/Vol-2414/paper17.pdf.

Chandrasekaran, M. K., Yasunaga, M., Radev, D., Freitag, D., & Kan, M.-Y.: Overview and results: CL-SciSumm Shared Task, . (2019). In Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2019) @ SIGIR 2019 (Pp. 2019). Paris: France.

Cohan, A., Ammar, W., Zuylen, M.V., & Cady, F. (2019). Structural scaffolds for citation intent classification in scientific Publications. In NAACL.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 2270–2282). Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.207. https://www.aclweb.org/anthology/2020.acl-main.207.

Davoodi, E., Madan, K., Gu, J. (2018). Clscisumm shared task: On the contribution of similarity measure and natural language processing features for citing problem. In BIRNDL@ SIGIR (Pp. 96–101).

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, Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423.

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, L. (2010). Context-aware citation recommendation. In Proceedings of the 19th international conference on World Wide Web, WWW '10, pp. 421–430. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1772690.1772734.

Hernandez-Alvarez, M., Soriano, J. M. G., & Martãnez-barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering, 23*(4), 561.

Jeong, C., Jang, S., Shin, H., Park, E., & Choi, S. (2019). A context-aware citation recommendation model with BERT and graph convolutional networks. *CoRR*.arXiv.org/abs/1903.06464.

Jeong, C., Jang, S., Shin, H., Park, E.L., Choi, S. (2020). A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, Pp. 1–16

Jha, R., Jbara, A. A., Qazvinian, V., & Radev, D. R. (2017). Nlp-driven citation analysis for scientometrics. *Natural Language Engineering, 23*(1), 93–130. https://doi.org/10.1017/S1351324915000443.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D. & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics 6*, 391–406 (2018). https://doi.org/10.1162/tacl_a_00028. https://www.aclweb.org/anthology/Q18-1028

Khalid, A., Khan, F. A., & Ahmed, I. (2017). Extracting reference text from citation contexts. *Cluster Computing, 21,* 1–18.

La Quatra, M., Cagliero, L., & Baralis, E. (2019). Poli2sum@cl-scisumm-19: Identify, classify, and summarize cited text spans by means of ensembles of supervised models. In M. K. Chandrasekaran, & P. Mayr (Eds.) Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2019) co-located with the 42nd international ACM SIGIR conference on research and development in information retrieval (SIGIR 2019), Paris, France, July 25, 2019, CEUR workshop proceedings (Vol. 2414, pp. 233–246). CEUR-WS.org. http://ceur-ws.org/Vol-2414/paper24.pdf

La Quatra, M., Cagliero, L., & Baralis, E. (2020). Exploiting pivot words to classify and summarize discourse facets of scientific papers. *Scientometrics, 125*, 1–19.

Li, L., Chi, J., Chen, M., Huang, Z., Zhu, Y., & Fu, X. (2018). Cist@clscisumm-18: Methods for computational linguistics scientific citation linkage, facet classification and summarization. In P. Mayr, M.K. Chandrasekaran, K. Jaidka (Eds.) Proceedings of the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2018) co-located with the 41st international ACM SIGIR conference on research and development in information retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018, *CEUR Workshop Proceedings* (Vol. 2132, pp. 84–95.) http://ceur-ws.org/. http://ceur-ws.org/Vol-2132/paper8.pdf.

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (pp. 74–81). Association for Computational Linguistics, Barcelona, Spain. https://www.aclweb.org/anthology/W04-1013.

Ma, S., Xu, J., & Zhang, C. (2018). Automatic identification of cited text spans: A multi-classifier approach over imbalanced dataset. *Scientometrics, 116*(2), 1303–1330. https://doi.org/10.1007/s11192-018-2754-2.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111–3119)

Moraes, L.F.T.D., Das, A., Karimi, S., & Verma, R.M. (2018). University of houston @ cl-scisumm 2018. In P. Mayr, M.K. Chandrasekaran, K. Jaidka (Eds.) Proceedings of the 3rd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2018) co-located with the 41st International ACM SIGIR conference on research and development in information retrieval (SIGIR 2018), Ann Arbor, USA, July 12, 2018, CEUR workshop proceedings (Vol. 2132, Pp. 142–149). http://ceur-ws.org/. http://ceur-ws.org/Vol-2132/paper14.pdf.

Nomoto, T. (2018). Resolving citation links with neural networks. *Frontiers in Research Metrics and Analytics, 3,* 31. https://doi.org/10.3389/frma.2018.00031.

Pagliardini, M., Gupta, P., & Jaggi, M. (2018) Unsupervised learning of sentence embeddings using compositional n-gram features. In Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies (Vol. 1, Pp. 528–540).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Pilehvar, M. T., & Camacho-Collados, J. (2020). Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies, 13*(4), 1–175.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3973–3983)

Ritchie, A., Robertson, & S., Teufel, S. (2008). Comparing citation contexts for information retrieval. In Proceedings of the 17th ACM conference on information and knowledge management, CIKM

'08 (Pp. 213–222). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1458082.1458113.

Ritchie, A., Robertson, S. & Teufel, S. (2008). Comparing citation contexts for information retrieval. In Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08 (Pp. 213–222). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1458082.1458113.

Ritchie, A., Teufel, S., & Robertson, S. (2006). How to find better index terms through citations. In Proceedings of the workshop on how can computational linguistics improve information retrieval?, CLIIR '06 (Pp. 25–32). Association for Computational Linguistics, USA

Ritchie, A., Teufel, S., & Robertson, S. (2008). Using terms from citations for IR: Some first results. In European conference on information retrieval, pp. 211–221. Springer

Ronzano, F., & Saggion, H. (2016). An empirical assessment of citation information in scientific summarization. In E. Métais, F. Meziane, M. Saraee, V. Sugumaran, & S. Vadera (Eds.), *Natural language processing and information systems* (pp. 318–325). Cham: Springer International Publishing.

Saier, T., & Färber, M. (2020). Unarxive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*. https://doi.org/10.1007/s11192-020-03382-z.

Wang, P., Li, S., Wang, T., Zhou, H., & Tang, J. (2018). Nudt@ clscisumm-18. In: BIRNDL@ SIGIR

Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A., Li, I., Friedman, D., & Radev, D. (2019). ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In Proceedings of AAAI 2019

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan & K., Radev, D.R. (2017). Graph-based neural multi-document summarization. In Proceedings of CoNLL 2017.

Yeh, J.Y., Hsu, T.Y., Tsai, C.J. & Cheng, P.C. (2017). Reference scope identification for citances by classification with text similarity measures. In Proceedings of the 6th international conference on software and computer applications, ICSCA '17 (p. 87–91). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3056662.3056692.

Yousif, A., Niu, Z., Chambua, J. & Khan, Z.Y. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. Neurocomputing **335**, 195 – 205. doihttps://doi.org/10.1016/j.neucom.2019.01.021. http://www.sciencedirect.com/science/article/pii/S0925231219300335

Zaki, M. J., & Meira, W., Jr. (2020). *Data mining and machine learning: fundamental concepts and algorithms* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108564175.

Zerva, C., Nghiem, M. Q., Nguyen, N. T., Ananiadou, S. et al. (2020). Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics*. https://doi.org/10.1007/s11192-020-03455-z.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1253.