Scientometrics
https://doi.org/10.1007/s11192-018-2944-y

Deep context of citations using machine-learning models in scholarly full-text articles

Saeed-Ul Hassan[1] · **Mubashir Imran**[1] · **Sehrish Iqbal**[1] · **Naif Radi Aljohani**[2] · **Raheel Nawaz**[3]

Received: 26 February 2018

Abstract
Information retrieval systems for scholarly literature rely heavily not only on text matching but on semantic- and context-based features. Readers nowadays are deeply interested in how important an article is, its purpose and how influential it is in follow-up research work. Numerous techniques to tap the power of machine learning and artificial intelligence have been developed to enhance retrieval of the most influential scientific literature. In this paper, we compare and improve on four existing state-of-the-art techniques designed to identify influential citations. We consider 450 citations from the Association for Computational Linguistics corpus, classified by experts as either important or unimportant, and further extract 64 features based on the methodology of four state-of-the-art techniques. We apply the Extra-Trees classifier to select 29 best features and apply the Random Forest and Support Vector Machine classifiers to all selected techniques. Using the Random Forest classifier, our supervised model improves on the state-of-the-art method by 11.25%, with 89% Precision-Recall area under the curve. Finally, we present our deep-learning model, the Long Short-Term Memory network, that uses all 64 features to distinguish important and unimportant citations with 92.57% accuracy.

Keywords Citation-context analysis · Deep learning · Influential citations · Machine learning

✉ Saeed-Ul Hassan
saeed-ul-hassan@itu.edu.pk

Naif Radi Aljohani
nraljohani@kau.edu.sa

Raheel Nawaz
r.nawaz@mmu.ac.uk

[1] Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

[2] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

[3] School of Computer Science, Manchester Metropolitan University, Manchester, UK

Published online: 29 October 2018

Introduction

We aim to investigate the problem of distinguishing cited work as either important or unimportant to the development of a scholarly publication. This is a vital task in qualitatively measuring the impact of publications in our growing scientific literature and in the behavioral analysis of scientific domains. The algorithms and techniques to approach a certain problem, as well as the writing style of the author (Hassan et al. 2017a, b), contribute greatly to making an article influential.

Traditionally, the absolute number of citations that an article receives is used to measure the impact of a scientific article (Abu-Jbara et al. 2013). Similarly, citation-based quantitative bibliometric metrics, such as Impact Factor (Garfield 2006), G-index (Egghe 2006), H-Index (Hirsch 2005, 2010a, b) and Scopus's source-normalized impact per paper (SNIP) (Waltman et al. 2013) are effective evaluators of the quantitative aspect of scientific articles. However, the question is whether all citations are as important as each other (Hassan et al. 2018a, b). The citation count is defined as the number of times a specific article has been referred to in preceding scientific literature (Lindsey 1989). However, the reference could concern the adoption of a particular method or be a mere acknowledgement of relevant background work. Valenzuela et al. (2015) argue that we cannot consider all citations as being of the same importance. While the number of citations of scientific publications can account for their quantitative impact (Borgman 1990; Luukkonen 1992), as a qualitative measure of impact not all citations can be considered equal.

Moravcsik and Murugesan (1975) found that about 40% of citations in their corpus of articles gave a perfunctory general acknowledgement. This explains the importance of a citation's context, since clearly a large number of citations are insignificant (Small and Greenlee 1980). Various annotation schemes have been devised to judge the importance of a cited work. In general, authors of scientific articles are most concerned with how useful a citation is in context. Recently, Teufel et al. (2006), Amjad et al. (2013), Valenzuela et al. (2015), Hassan et al. (2017a, b) and Hassan et al. (2018a, b) present various models to identify the importance or unimportance of scientific articles as they are referred to in some works. They identify various features of the citation's context (the text and sections surrounding a citation). Our contributions in this research direction are as follows:

1. We present a supervised machine-learning classification model named Hassan_29 to select the best-performing features using the Extra-Trees classifier, which improves on the state-of-the-art classifier by Valenzuela et al. (2015) by 11.25%, with 89% under the Precision-Recall (PR) curve, using the Random Forest (RF) classifier.
2. We present the Long Short-Term Memory (LSTM)-based deep-learning model to distinguish between important and unimportant citations, and this outperforms traditional machine-learning models, achieving an accuracy of 92.5%.

Related work

Conventionally, citation analysis has been used to measure the quality of an article in the scientific literature, hence the tracking of citations plays a vital role. It has been argued that not all citations are equal, therefore a classification is needed to distinguish the important

Table 1 Literature review summary

Type	References
Citation Context	Nanba and Okumura (1999), Pham and Hoffmann (2003), Nakov and Okumura (2004), Bertin and Atanassova (2018), Cohan and Goharian (2017), Taskin and Al (2018) and Peritz (1983)
Citation Classification	Moravcsik and Murugesan (1975), Teufel et al. (2006), Hassan et al. (2017a, b), Garfield (1965), Chubin et al. (1975), Oppenheim and Renn (1978), Frost (1979), Finney (1979), Garzone and Mercer (2000), Conrad and Dabney (2001), Agarwal et al. (2010), Xu et al. (2013), Ding et al. (2014) and Pride and Knoth (2017)
Citation Sentiment	Amjad et al. (2013), Athar (2011), Pang and Lee (2008), Zhang and Koppaka (2007), Hou et al. (2011) and Balaban (2012)

from the unimportant. Further, we present a brief review on related studies, see Table 1 for a quick reference that summarizes literature review into the following themes: Citation Context, Citation Classification and Citation Sentiments.

Brief review of citation context

Nanba and Okumura (1999) used cue phrases to classify a citation type as basic, comparison or 'other'. These cue phrases around a citation were selected manually, and the overall system achieved an accuracy of 83%. Pham and Hoffmann (2003) proposed a new system to reduce the time spent in manually listing the cue words. The system consists of ripple-down rules. The rules are simple patterns comprising a random number of words and gaps between them. This system classifies citations into basic, support, limitation and comparison. The system performs better than that of Nanba and Okumura (1999) and achieves good accuracy. Nakov et al. (2004) recognized the use of context in text summarization. They provided the information to summarize literature using important facts, for example the text around the citation.

Bertin and Atanassova (2018) considered multiple in-text references and their position in an article. For this purpose, they used a dataset of 80,000 research articles. They analysed two characteristics: the position of Multiple In-text References (MIR) and the total number of references that make up a MIR. Cohan and Goharian (2017) first addressed the problem of inaccurate citation-context extraction, suggesting a new method for making an automatic summary of research articles by using the context of citations. They used a dataset from the biomedical and computational linguistics domain. Peritz (1983) introduced a method for labelling citations for the assessment of both quality and context. She stated that existing classification systems are inappropriate as the role of citation varies between one discipline and another, and proposed a new scheme of eight categories. She observed that the negational (or disagreement) class occurs most frequently in the literature.

Brief Review of Citation Classification

Garfield (1965) was among the founders of bibliometric methods and a pioneer in the field of scientometrics, proposing a citation classification scheme that acknowledges that

authors might have contrasting perspectives when citing publications. He speculated on the various reasons why an author might cite an article, as shown in Table 2. Moravcsik and Murugesan (1975) classified citations into: conceptual vs operational; evolutionary or juxtapositional; organic or perfunctory; and confirmative and. Chubin and Moitra (1975) adapted the scheme of Moravcsik and Murugesan (1975) by making absolute categories manually. Their taxonomy uses the classification categories of affirmative, negational, basic and supplementary. Their results also show that the one that occurs most frequently in the literature is negational citation.

Oppenheim and Renn (1978) proposed a unique scheme to classify citations in the physical sciences, explaining why older articles are cited more than newer ones. Their study revealed that 40% of citations are for historical reasons, and only the remaining 60% are citations of previous articles in any active sense. Frost (1979) proposed a scheme to study the nature of citations to criticize and to handle citations and quotations in the principal literature on German literary works. Most of her categories correlate to those of Weinstock, and her main development lies in recommending two new categories to discriminate between the humanities and scientific works. Finney (1979) was the first to introduce an automated citation classifier. She introduced a seven-category scheme to classify citations in the medical literature, suggesting that the classification of scientific literature should be based on the cue words around a citation and on the location of a citation in the article. Her system fails, due to its small number of citation categories.

Garzone and Mercer (2000) used 200 manually selected rules to expand the classification to 35 categories, consisting of the following generic types: affirmational; negational; tentative; assumptive; developmental; methodological; future research; contrastive; and, finally, citations that utilize the conceptual. Their model achieved fair performance on six unseen articles. Conrad and Dabney (2001) proposed a system to distinguish citations that are manually checked by professional editors. The system comprises 20 hand-coded rules to identify distinct patterns. The system grants various forms of words and the presence of synonyms and gaps, but the explicit rules for language are not shown. Testing the system on an unseen dataset achieved a precision of 9.15% and recall of 59.09%, which indicates that the generation of rules achieves high precision.

Teufel et al. (2006) suggested a method to classify the citation function automatically by using several shallow and linguistically inspired features: a finite grammar using strings with part-of-speech-based recognition of actions. These features are used in association with their location and verb tense. The authors adopted the supervised classification model IBk, with 10-fold cross-validation, and achieved an accuracy of 79% and an F1 measure of 68%. Agarwal et al. (2010) also proposed an automated model for the classification of citations. They used the annotated corpus of full-text biomedical articles and the supervised classification techniques of Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB). The features that they used are unigrams and bigrams, and the rank of a feature was defined by manual information. They achieved an F-measure of 76.5%, and the SVM model outperformed the MNB model.

Xu et al. (2013) proposed a citation classification using three classes: functional; ambiguous; and perfunctory. They used distinct features for this classification, such as cue patterns, positional features, network-based features and structural features. These measure the relationship between the author and the article. Ding et al. (2014) proposed a method to identify important citations in scholarly big literature. Citations mentioned for the purpose of using or extending the work are considered to be important citations. The authors divided citations into related work, comparison, using the work and extending the work. They used the supervised classification models SVM and RF, using a three-fold class validation, and achieved overall accuracy of 80% with both. Pride and Knoth (2017) worked

Table 2 Reasons for citing an article

S. no	Reason to cite an article
1	To pay homage
2	To give credit
3	To identify methodology and equipment
4	To provide background studies
5	To correct own work
6	To correct the work of others
7	To criticize
8	To substantiate a claim
9	To give notification of a forthcoming work
10	To provide a lead to poorly indexed or uncited work
11	To authenticate data and classes of fact
12	To identify the original publication in which a concept is explained
13	To identify the original publication or other work defining an eponymous idea or term
14	To disclaim the work and concepts of other
15	To spread the claims of others regarding the property

on the classification of citations on the basis of their individual importance. Their results confirmed that multiple in-text references are highly predictive of influence.

More recently, Hassan et al. (2017a, b) extended the work of Valenzuela et al. (2015) by exploring novel features to classify citations as either important or unimportant. Their new features perform the best of the five supervised classification techniques of SVM, K-Nearest Neighbor (KNN), Naïve Bayes, Decision Tree and RF. Their RF model outperforms Valenzuela's model, achieving an overall accuracy of 84%.

Brief review of citation sentiment

Athar (2011) worked on the problem of determining positive and negative sentiments in the citations in scientific articles using the appended category of objective, along with a handful of features for classification. Pang and Lee (2008) worked on citation classification using sentiment analysis and opinion mining. This type of citation analysis concludes that an author cites a particular article either for support or to determine its weaknesses. Amjad et al. (2013) extended the work of Teufel et al. (2006) by suggesting a mechanism to identify the citation context, classify the citations and perform sentiment analysis. Different context-level and polarity-level features were needed for this task. They used the supervised classification model of SVM with 10-fold cross-validation, achieving an accuracy of 81.4%. Their results show that adding context to the citation improves the results and that two-way classification outperforms other methods. Zhang and Koppaka (2007) suggested a method to differentiate legal citations, because the citations in any single specific work will target a distinct case or issue. They built a network of citations, bearing in mind the legal principles. Each citation focuses on single legal case, so the number of cases that the researcher has to manage is reduced.

Table 3 Citation labelling

Label	Label tag	Description
0	Incidental/unimportant	Indicates unimportant citations
1	Important	Indicates important citations

Table 4 Annotated dataset

Annotator	Article ID	Cited by	Citation frequency
A	A97-1011	A00-2017	1
A	C00-1072	P02-1058	2
B	…	…	…

Hou et al. (2011) introduced a new scheme of counting citations in text. They divided the citations into two groups namely: closely related references and less related references. Of the total 651 articles examined in the fields of Biochemistry and Molecular Biology and Genetics & Heredity in the Web of Science, the authors showed that on average the closely related references appeared 3.35 times in full-text, compared to the less related references with 1.88 times only. Balaban (2012) proposed a technique to give more weight to citations from renowned authors. He also suggested that a citation of an article that had been published in a journal with a low impact factor should be regarded as more important. Consequently, the worth of a citation is inversely related to the impact factor of the journal in which the cited article was published.

Data and Method

We used a manually annotated and publicly available dataset from the Association for Computational Linguistics (ACL) (Valenzuela et al. 2015). There are 20,527 articles available in this ACL anthology.[1] These contain a total 106,509 citations, of which 450 were randomly selected then annotated as incidental or important (0/1), as shown in Table 3. Note that we refer these citation as tagged citations from here on. As shown in Table 4 as *Citation frequency*, each citing article may have tagged citations that occur one or more times. The tagged citations were further verified by a group of two experts, who were provided with the full text of the articles. The inter-annotator agreement was 93.9%. Note that, in this dataset, the annotators considered 14.6% of the citations to be important and 85.4% as incidental (unimportant).

Data Extraction and Pre-processing

We used the following pre-processing steps to extract citations and features from the dataset of articles: (a) we appended the article's given ID (e.g. P05-1044) with the anthology's

[1] http://allenai.org/data.html.

Table 5 Extracted citation context

Article	Cited by	Citation context window
A00-1043	C00-2140	"We shorten the output of the summarizer to a telegraphic style"; that way, more information can be included in a summary of k words (or n bytes)
		"Since we only use shallow methods for textual analysis that do not generate a dependency structure, we cannot use complex methods for text reduction as described, e.g., in (*Jing, 2000*)"
		"Our method simply excludes words occurring in the stop-list from the summary, except for some highly informative words such as 'I' or 'not'"
		"Since we want to enable interactive summarization which allows a user to browse through a dialogue quickly to search for information he is interested in, we have integrated our summarization system into a JAVA based graphical user interface ("Meeting Browser") (Bett et al. 2000)"

URL (www.aclweb.org/anthology/[article's ID here]) to retrieve the full-text article from the ACL anthology in pdf format; (b) we used Poppler's pdf-to-text (http://poppler.freedesktop.org) to extract the text from the pdf file of each research article; (c) furthermore, we used regular expressions to identify the occurrence of a particular (tagged) citation in the text; (d) we used the Stanford Parser[2] to parse the citation text and obtain the text surrounding a particular citation (i.e. citation context window of four sentences), as depicted in Table 5. A citation context window consists of one sentence before the tagged citation and two sentences after (Abu-Jbara et al. 2013); the above methodology also identifies citations (other than tagged citations) occurring within the citation context window; (e) on these sets of sentences (citation windows), we used OpenNLP[3] library, for parts of speech (POS) tagging, as identified in Table 6; (f) finally, to identify the sections in which the citation occurred, ParsCit[4] was used identify the section of the tagged citation. Note that if the citation frequency (Table 4) in the citing article is more than one, logical OR is taken for all binary features and the mean is taken for continuous features.

Finally, creating citation context windows and tokenizing them helped us to extract window features (i.e. other references, multiple references, reference count, is separate, etc.). The POS tagging helped to extract various features, such as the demonstrative determiner, closest verb/adjective/adverb, contain 1st/3rd person pronoun or contain closest noun phrase, and so on.

Deployed models

In this section, we describe the data extraction and machine-learning approaches deployed by Amjad et al. (2013), Valenzuela et al. (2015), Teufel et al. (2006) and Hassan et al. (2017). From here on, we refer to each as described in Table 7.

[2] https://nlp.stanford.edu/software/lex-parser.shtml.
[3] http://opennlp.apache.org/.
[4] http://parscit.comp.nus.edu.sg.

Table 6 POS tagging of citation context provided in Table 5

Tag	Token	Section
Noun	Output, summarizer, style, way, information, summary, words, bytes, methods, analysis, dependency, structure, text, reduction, stop-list, user, browse, dialogue, search, JAVA, interface, browser	*Experiment*
Pronoun	We, our, he, I	
Verb	Shorten, can, included, use, do, generate, described, excludes, occurring, want, enable, allows, interested, integrated, based, meeting	
Adverb	Only, simply, highly, quickly	
Adjective	More, shallow, textual, complex, informative, interactive, summarization, graphical	
Determiners	This, that, these, those (predefined)	

Table 7 Alias of approaches described in articles

Referring article	Venue	Approach name
Teufel et al. (2006)	EMNLP (2006)	Teufel model
Amjad et al. (2013)	NAACL (2013)	Amjad model
Valenzuela et al. (2015)	AAAI (2016)	Valenzuela model
Hassan et al. (2017)	JCDL (2017)	Hassan model

Table 8 Teufel's features

Feature	ID	Description
Weak	T-F1	Citing article mention weakness of cited article
CoCoGM	T-F2	Citing article compare/contrast methods or goals with cited article
CoCo-	T-F3	Citing article work is superior to cited article
CoCoRO	T-F4	Comparison of 2 cited articles
CoCoXY	T-F5	Contrast between cited articles
PBas	T-F6	Author uses cited work as base
PUse	T-F7	Author uses tools/algorithm of cited article
PModi	T-F8	Author modifies cited work
PMod	T-F9	Citation used to motivate current work
PSim	T-F10	Similarity of cited and citing work
PSup	T-F11	Citing and cited work are compatible
Neut	T-F12	Neutral description of cited work

Teufel model

In their work, Teufel et al. (2006) extracted 12 basic features to describe the various capacities in which a citation may be used. All features, along with an identifier, are presented in Table 8. Each was divided into four categories: weakness; comparison; sentiments; and neutral. Further, they classified them as weak, positive or neutral. They achieved an accuracy of 83% using IBk ($k=3$) classifier, using WEKA. Teufel created four additional

Table 9 Amjad's features

Feature	ID	Description
Demonstrative determiner	A-F1	Citation context contains demonstrative determiner
Conjunctive adverb	A-F2	Citation context contains conjunctive adverb
Position	A-F3	Position of citing sentence
Contains closest noun phrase	A-F4	Citation context contains closest noun phrase
Other reference	A-F5	Citation context contains reference other than target
Mention of target	A-F6	Citation context contains the mention of target reference
Multiple references	A-F7	Target citation sentence contains multiple references
Criticizing	A-F8	Citing article mentions weakness/strengths of cited article
Comparison	A-F9	Citing article compares/contrasts work with cited article
Use	A-F10	Citing article uses the work of cited article
Substantiating	A-F11	Citing article is similar/supports the cited work
Basic	A-F12	Citing article uses cited article as a starting point
Neutral	A-F13	Citing article
Reference count	A-F14	Number of references in context
Is separate	A-F15	Citation occurs separately
Closest verb/adjective/adverb	A-F16	Distance of closest verb, adjective or adverb
Self-citation	A-F17	Citation is self-citation
Contains 1st/3rd person pronoun	A-F18	Context contains 1st/3rd person pronoun
Negation cue	A-F19	Context contains negation cue
Speculation cue	A-F20	Context contains speculation cue
Subjectivity cue	A-F21	Context contains subjectivity cue
Contrary expression	A-F22	Context contains contrary expression
Section	A-F23	Section of citation

features (Negative, Positive, Contrast and Neutral+) by combining the features mentioned below.

Amjad model

Amjad et al. (2013) applied reference tagging, reference grouping and non-syntactic reference removal to extract three sets of features. These are defined as context identification, purpose of citation and polarity. A total of 22 features were extracted, as presented in Table 9. The authors applied SVM (kernel = linear, $c = 1.0$) to context-identification features and achieved a precision of 92% and a recall of 76.4%.

Among the contextual features, they noted that lexical features were generally more important than structural features. In their classification of a citation's purpose, they achieved an accuracy of 70.5%. They noted from their results that authors first make a citation by using a neutral sentence, then follow it with a critical one. They computed and compared Pearson correlation coefficients among the polarity and purpose features. They found a high correlation between Use and Basis, and concluded that when authors present new technology the algorithms and corpora used by that scientific research start a trend, and thus generate more citations.

Table 10 Valenzuela's features

Feature	Feature ID	Description
F1	V-F1	Direct citations
F2	V-F2	Direct citations per section
F3	V-F3	Indirect citations
F4	V-F4	Author overlap
F5	V-F5	Is useful
F6	V-F6	In figure/table
F7	V-F7	Inverse no. of references
F8	V-F8	All citations
F9	V-F9	Abstract similarity
F10	V-F10	Page rank
F11	V-F11	Total citing articles
F12	V-F12	Domain of the cited article

Valenzuela model

Valenzuela et al.'s (2015) extracted features mostly relate to the nature of the citation and the section in which it appears. A description of these features is provided in Table 10. The authors constructed a supervised classification model with SVM (kernel = RBF) and RF. Both classifiers obtained an encouraging 80% of the area under the curve (AUC). They incorporated their model into a search engine for scientific literature.

Hassan model

Hassan et al. (2017a, b) extended the work of Valenzuela et al. (2015) and presented 13 features. These are categorized into three groups: context-based features; cue word-based features; and textual features. They constructed a model with five classifiers, namely RF, SVM, KNN, Decision Tree and Naïve Bayes. RF was their best-performing classifier, with an encouraging AUC of 91%. This showed that the RF classifier discriminated very well between important and unimportant citations. The authors applied the Extra-Trees Classifier to compare the performance of individual features and showed that Feature H-F13 (Abstract and text similarity), as presented in Table 11, is more informative than the others. It is followed by H-F1 (Total citations received by reference) and H-F11 and H-F12 (Cue words for using and extending existing work). Note that that they merged H-F9 and H-F10 as H-F9, and merged H-F11 and H-F12 as H-F12, since they found a significant overlap of keywords among these features.

Proposed models

We propose two supervised traditional machine-learning models, namely SVM and RF, and an LSTM-based deep-learning model to address the problem of citation classification.

Table 11 Hassan's features

Feature	ID	Description
F1	H-F1	Total citations received by reference
F2	H-F2	Total citations
F3	H-F3	Citations in introduction section
F4	H-F4	Citations in literature review section
F5	H-F5	Citations in method section
F6	H-F6	Citations in experiment section
F7	H-F7	Citations in discussion section
F8	H-F8	Citations in conclusion section
F9	H-F9	Cue words for related work
F10	H-F10	Cue words for comparative citations
F11	H-F11	Cue words for using and extending existing work
F12	H-F12	Cue words for extending existing work
F13	H-F13	Abstract and text similarity
F14	H-F14	Author overlap

Table 12 Classifier parameter settings for each model

Model	Tuned parameter				
	SVM			RF	
	kernel	ɣ	C	Estimators	max_features
Teufel	RBF	0.1	1	100	5
Amjad	RBF	0.1	1	100	5
Valenzuela	RBF	0.1	1	100	5
Hassan	RBF	2	0.5	100	5
Hassan_29	RBF	0.1	1	100	5

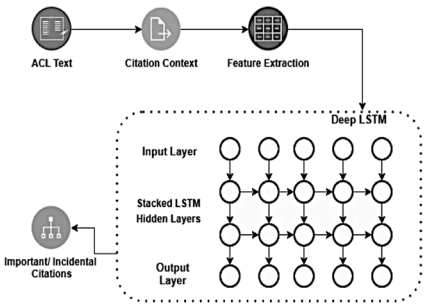
Supervised model

For the purpose of comparison, we employed supervised classification techniques by applying the SVM (Auria and Moro 2008) and RF (Breiman 2001) classifiers to the feature set presented by each model. SVM finds the optimal boundaries of the outputs by transforming data using a specific kernel. Here, we applied a non-linear Radial Basis Kernel (RBF) for transformation(Cao et al. 2008). The RBF function is provided in Eq. 1.

$$k(x, z) = e^{-\gamma} \parallel x - z \parallel 2, \gamma > 0 \tag{1}$$

Here $e^{-\gamma}$ is a constant, while x and z represent vectors in some feature space. RF is a supervised machine-learning algorithm that, as its full name suggests, creates a forest of classification trees and splits the feature nodes randomly. We computed the precision, recall, F1-score and Precision-Recall (PR) curve to compare the performance of each model on same dataset. The parameter settings for both SVM and RF are presented in Table 12.

These settings are tuned parameters for each set of features and are analogous to the parameters used by Abu-Jbara et al. (2013) and Hassan et al. (2017). We selected SVM and RF because three of the compared adopted models—by Abu-Jbara et al. (2013), Valenzuela et al. (2015) and Hassan et al. (2017a, b)—outperformed the other classifiers. To extract their best features, we employed the Extra-Trees classifier (Geurts et al.



Require: $input\ Result_{last}\ Cell_{last}\ w$**(input weight matrix)** u **(weight matrix of the last result)** b**(bias vector)** r **(projection matrix)**
Ensure: $Cell_{Value}\ Result$

while *When having input* **do**

$temp_{input} = \dot{}(Result_{last} * u_{input}) + \dot{}(input * w_{input}) + b$

$temp_{input} = \dot{}(Result_{last} * u_{output}) + \dot{}(input * w_{output}) + b$

$temp_{cell} = \dot{}(Result_{last} * u_{cell}) + \dot{}(input * w_{cell}) + b$

$input_{gate} = sigmoid(temp_{input})$

$output_{gate} = sigmoid(temp_{output})$

$forget_{gate} = 1 - inpu\ t_{gate}$

$Cel\ _{Value} = actiVate(temp_{cell}) * input_{gate} + Cell_{last} * forget_{gate}$

$Result = output_{gate} * actiVate(cell_{Value})$

$Cell_{Value} = \dot{}(Result, r)$

end while

Fig. 1 LSTM-based deep-learning model and pseudocode

2006), also known as the 'Extremely randomized trees classifier', to split the complete selection of data at each step and randomly select a decision boundary. For the final feature selection for our model, we selected all 29 features that had an Extra-Trees classifier score of more than 0.01. We named the machine-learning model 'Hassan_29' (see Appendix, Table 13).

Deep-learning model

In recent years, deep-learning neural networks have overthrown conventional machine-learning algorithms in both supervised and unsupervised tasks (Schmidhuber 2015). The deep-learning classification model can be thought of comprising layers of non-linear units that perform transformation and feature extraction tasks (Di Ciaccio and Giorgi 2015). A Deep Neural Network (DNN) consists of a number of hidden layers, on which each utilizes the output of the layer before as its input. An improved variation of Neural Networks is a Recurrent Neural Network (RNN), with a short-term memory to retain the contextual information from previous results. Figure 1 shows our LSTM-based deep-learning model and its pseudocode.

We used the Keras implementation of the LSTM network (Hochreiter and Schmidhuber 1997) to solve our classification problem. LSTM is a variant of an RNN that uses the short-term memory of an RNN neuron and makes it last longer. This is accomplished through a special module in LSTM that controls the information to be used. Our implementation of the Keras LSTM model uses TensorFlow at the backend (Abadi and TensorFlow 2016). It consists of six layers, each dividing the dimensions (neurons) of the previous layers in two.

Finally, to convert the weighted results of each neuron into output and to introduce non-linearity in our network, we applied a sigmoid activation function at each layer. A sigmoid function is suitable here because most of our features were between 0 and 1, or normalized between 0 and 1. Equation 2 represents the output of a neuron (z), where w represents the weights and x represents the inputs. This output is fed to Eq. 3, where we matched the weight to the activation σ(z).

$$z = \sum_{i=1}^{m} w_i x_i + \text{bias} \tag{2}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

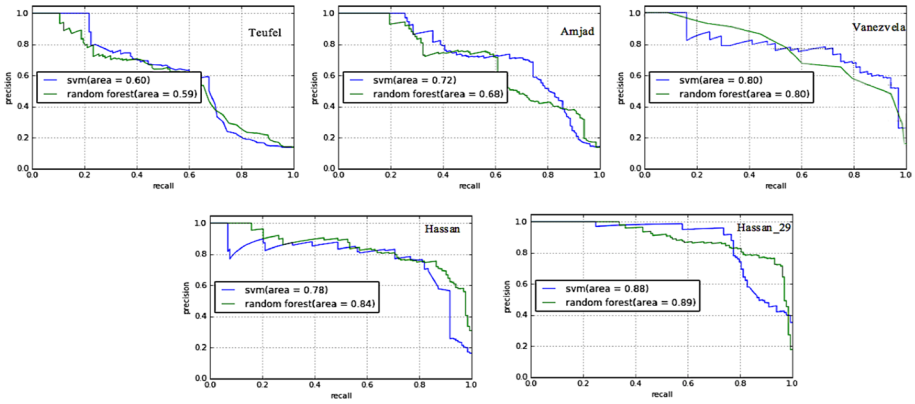


Fig. 2 PR curve for SVM and RF classifier across the deployed models, using 10-fold cross-validation

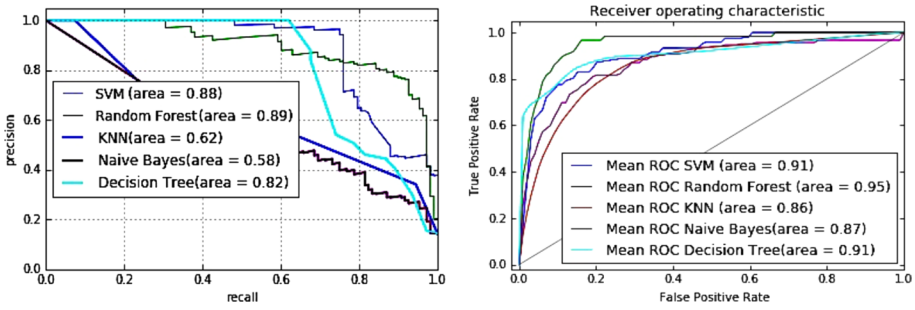


Fig. 3 PR and ROC curve for SVM, KNN, Decision Tree and Naïve Bayes classifier on 29 influential features, using 10-fold cross-validation

Results and discussion

This section describes our results from a series of experiments. We divided our data into two parts, consisting of training then testing the data using three-fold cross-validation. We trained our model on training data and then evaluated it on testing data. Our aim in these experiments was to compare four state-of-the-art techniques and compare them with a newly proposed model for the classification of citations as important or unimportant. For this purpose, we used PR curves, using five different supervised classifiers and comparing the results with the deep-learning model. The classifiers that we used in our experiment are Naïve Bayes, SVM, RF, Decision Tree and KNN.

Table 12 shows a summary of the classification results of all five models, including our machine-learning based proposed model, Hassan_29. We applied SVM and RF on the features extracted through methodologies, and the parameter settings for each model are shown in Table 11. All models performed fairly well in predicting the important and unimportant citations. According to our results, RF outperforms SVM in all models. In addition to this summary of the classification results, we show the PR curve of each model in Fig. 2. This shows that our model has greater precision than other models in terms of recall, with an f-measure reaching 0.91 for the RF models.

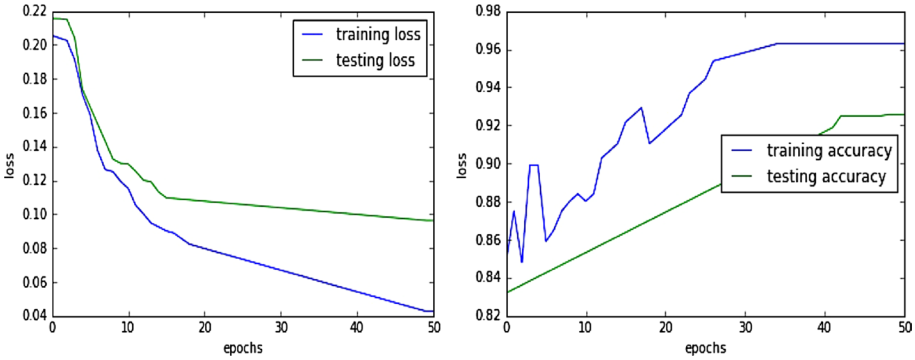


Fig. 4 Training losses and accuracy through 50 epochs

The results show that, with Valenzuela, Hassan and Hassan_29, RF outperforms SVM. The reasons behind these results are that these models have of a mixture of continuous and numeric features, and that the citation features contain outliers. In such conditions, RF performs well. In addition to RF and SVM, we also used KNN, Decision Tree and Naïve Bayes on the 29 influential features by using three-fold cross-validation techniques and observed PR curve. As can be seen in Fig. 3 (left side), RF still outperforms all the other classifiers, with a PR = 0.89. SVM also gives a better performance, with a PR = 0.88, while Naïve Bayes performed the worst of all, with a PR = 0.58.

Furthermore, we used ROC curves to evaluate our model and show how well it differentiates the important from the unimportant citations. We used a three-fold cross-validation technique to train the classifier. Figure 3 (right side) shows the ROC curves for all five models. We found that RF beats all the other classifiers, with an ROC = 0.97. RF achieved better performance with an ROC = 0.95 and SVM also performed well, with an ROC = 0.91. We concluded that, overall, the RF classifier is a better predictor than the others. Overall, RF has better results due to its ability to classify effectivity even when there is deviance in the data. The Naïve Bayes classifier performs worst, because the data size is small and the assumptions on which Naïve Bayes is based appear not to hold with the experimental dataset. It cannot learn the interactions between the features and is not robust in learning, hence, resulting in poor performance.

Finally, we deployed our deep-learning model with the Keras DNN. Our model consists of an input layer with 52 units and five hidden layers of 26, 13, 7, 3 and 1 units respectively. Each layer uses a sigmoid as the activation function. Testing and training sets are in a ratio of 9:1 and were randomly picked numerous times. Our models achieved an average accuracy of 92.57%, which is very good, considering the size of the inputs. Figure 4 shows the learning rate at testing, and the training losses and accuracy through 50 epochs. Our model showed significant improvement up to 30 epochs, then the training and testing accuracy levelled off. Overall, our deep-learning model outperformed traditional machine-learning models with an accuracy of 92.57%. However, given the small dataset, the improvement on traditional machine learning and deep learning is not clearly evident.

Conclusion

Our work is the first attempt to compare the state-of-the-art models for classifying the importance of a citation using the same dataset. We have shown that our machine-learning model, with top 29 features, outperforms all existing state-of-the-art models. In addition, our deep-learning based LSTM model, with all 64 features, does exceptionally well in identifying the importance of a citation for a given article, with an accuracy of more than 92%.

Citation-based indices are a major tool used by research administrators for academic assessment. The most renowned indices such as h-index, impact factor, source-normalized impact per paper, and so on, are quantitative in nature and give no credit to the importance of the context of a citation within an article. Moreover, these indices use absolute citation counts, which may fail to distinguish the significance of an important work. Therefore, bibliometric indices that measure the impact of a scientific article on the basis of its context are of paramount importance. We believe that identifying the context in which an article may be vital and prove to be a more informative measure of its impact.

Our approach can be used to enhance state-of-the-art specialist information extraction techniques, such as the meta-knowledge annotation scheme of Thompson et al. (2011) and the hypothesis or new-knowledge detection scheme of Shardlow et al. (2018). For example, the incorporation of the relative importance of a citation can help to refine the knowledge type/category encapsulated in a statement.

Another key application of our work would be in establishing the 'global' and 'local' importance of a research article. For example, rather than scoring articles by the total number of citations that they receive, more sophisticated schemes can be developed to establish the importance of individual citations within an article. These 'local' scores for a cited article can be collected for all citations of the article in question, and a 'global' importance score synthesised. Using such global importance scores, one can establish the 'actual/qualitative' significance of an article.

A potential limitation of our work lies in the definitions of 'important' and 'incidental' (unimportant) citations. This study adopted the definitions that came with the standard dataset, yet these may not necessarily be accurate. Another limitation of our work is the difficulty to adapt it to scholarly big data, since some of the proposed features are manually computed. When scaling up this study to larger datasets, such features could be extracted using cue word based approaches. For example, conjunctive adverbs (A-F2) can be obtained by specifying cue words and parsing the sentence to compute their occurrence. Similarly, for features such as H-F1, automated crawlers can be built to extract feature data from the web and, for features such as H-F13, sentiment-based models can be built to check and validate the similarity score.

Overall, our proposed technique contributes to the emerging field of bibliometric-enhanced information retrieval by increasing the query search capabilities of search engines and semantic search approaches on Web 2.0 (De Vocht et al. 2017; Jiang and Yang 2018). Last but not least, this work can help improve citation-based full-text summarization techniques.

Note that the data and code to reproduce all the analysis presented this paper may be downloaded from the following URL: https://github.com/slab-itu/imp_citations.

Appendix

See Table 13.

Table 13 Scores of the top 29 features selected using Extra-Trees classifier

Feature name	Feature ID	Classification score
Method	H-F5	0.067301
Citation in article	H-F2	0.055028
PUse	T-F7	0.046484
Author_overlap	V-F4	0.039698
Section	A-F23	0.039119
abs_cite_similarity	V-F9	0.037921
Contrast	T-F13	0.037383
Normalized_cites_per_year	V-F8	0.033105
Use	A-F10	0.031846
Total citation	V-F1	0.030411
Related_work	H-F9	0.028996
Reference count	A-F14	0.028723
Closest verb/adverb/adjective	A-F16	0.028289
Substantiating	A-F11	0.028132
Compare	H-F10	0.026866
Using	H-F11	0.025311
Demonstrative determiner	A-F1	0.023467
Contain closest noun phrase	A-F4	0.021491
Basics	A-F12	0.020505
Introduction	H-F3	0.020111
Multiple references	A-F7	0.019426
Conjunctive adverb	A-F2	0.018472
PSim	T-F10	0.018177
PSup	T-F11	0.017775
Neutral+	T-F15	0.016902
PBas	T-F6	0.016614
Is separate	A-F15	0.016511
1st/3rd personal pronoun	A-F18	0.016458
Speculation	A-F20	0.015999

References

Abadi, M., & TensorFlow, A. A. B. P. (2016). Large-scale machine learning on heterogeneous distributed systems. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16), Savannah, GA, USA* (pp. 265–283).

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 596–606).

Agarwal, S., Choubey, L., & Yu, H. (2010). Automatically classifying the role of citations in biomedical articles. In *AMIA Annual Symposium Proceedings* (Vol. 2010, p. 11). American Medical Informatics Association.

Athar, A. (2011, June). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session* (pp. 81–87). Association for Computational Linguistics.

Auria, L., & Moro, R. A. (2008). *Support vector machines (SVM) as a technique for solvency analysis*. Technical report, Deutsche Bundesbank, Hannover; German Institute for Economic Research, Berlin. (2007)

Balaban, A. T. (2012). Positive and negative aspects of citation indices and journal impact factors. *Scientometrics, 92*(2), 241–247.

Bertin, M., & Atanassova, I. (2018). The context of multiple in-text references and their signification. *International Journal on Digital Libraries*, *19*(2-3), 287-303.

Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., & Waibel, A. (2000). Multimodal meeting tracker. In *Content-Based Multimedia Information Access* (Vol. 1, pp. 32–45).

Borgman, C. L. (1990). *Scholarly communication and bibliometrics*. Newbury Park: Sage Publications.

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

Cao, H., Naito, T., & Ninomiya, Y. (2008, October). Approximate RBF kernel SVM and its applications in pedestrian classification. In *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423–441.

Cohan, A., & Goharian, N. (2017). Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, *19*(2–3), 287-303.

Conrad, J. G., & Dabney, D. P. (2001, October). Automatic recognition of distinguishing negative indirect history language in judicial opinions. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 287–294). ACM.

De Vocht, L., Softic, S., Verborgh, R., Mannens, E., & Ebner, M. (2017). Social semantic search: a case study on web 2.0 for science. *International Journal on Semantic Web and Information Systems, 13*(4), 155–180.

Di Ciaccio, A., & Giorgi, G. M. (2015). Deep learning for supervised classification. *Rivista Italiana di Economia Demografia e Statistica*, *69*(2), 2–10.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology, 65*(9), 1820–1833.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Finney, B. (1979). *The reference characteristics of scientific texts*. Doctoral dissertation, City University (London, England).

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *The Library Quarterly, 49*(4), 399–414.

Garfield, E. (1965, December). Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 269, pp. 189–192). Washington, DC: National Bureau of Standards, Miscellaneous Publication 269.

Garfield, E. (2006). The history and meaning of the journal impact factor. *The Journal of the American Medical Association, 295*(1), 90–93.

Garzone, M., & Mercer, R. (2000). Towards an automated citation classifier. In *Conference of theCanadian Society for Computational Studies of Intelligence* (pp. 337-346). Springer, Berlin.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning, 63*(1), 3–42.

Hassan, S. U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, (pp. 1–8). IEEE.

Hassan, S. U., Imran, M., Iftikhar, T., Safder, I., & Shabbir, M. (2017). Deep stylometry and lexical & syntactic features based author attribution on PLoS digital repository. In *International Conference on Asian Digital Libraries* (pp. 119–127). Springer, Cham.

Hassan, S. U., Iqbal, S., Imran, M., Aljohani, N. R., & Nawaz, R. (2018). Mining the context of citations in scientific publications. In *International Conference on Asian Digital Libraries* (in-press). Springer, Cham.

Hassan, S. U., Safder, I., Akram, A., & Kamiran, F. (2018b). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics, 116*(2), 973–996.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569.

Hirsch, J. E. (2010a). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics, 85*(3), 741–754.

Hirsch, J. E. (2010b). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics, 85*(3), 741–754.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
Hoffmann, A., & Pham, S. B. (2003, October). Towards topic-based summarization for interactive document viewing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 28–35). ACM.
Hou, W. R., Li, M., & Niu, D. K. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays, 33*(10), 724–727.
Jiang, Y., & Yang, M. (2018). Semantic search exploiting formal concept analysis, rough sets, and Wikipedia. *International Journal on Semantic Web and Information Systems (IJSWIS), 14*(3), 99–119.
Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. *Scientometrics, 15*(3–4), 189–203.
Luukkonen, T. (1992). Is scientists' publishing behaviour rewards eeking? *Scientometrics, 24*(2), 297–319.
Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86–92.
Nakov, P. I., Schwartz, A. S., & Hearst, M. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR* (Vol. 4, pp. 81–88).
Nanba, H., & Okumura, M. (1999, July). Towards multi-paper summarization using reference information. In *IJCAI* (Vol. 99, pp. 926-931).
Oppenheim, C., & Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the Association for Information Science and Technology, 29*(5), 225–231.
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1–135.
Peritz, B. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics, 5*(5), 303–312.
Pride, D., & Knoth, P. (2017, September). Incidental or influential? Challenges in automatically detecting citation importance using publication full texts. In *International conference on theory and practice of digital Libraries* (pp. 572–578). Springer, Cham.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85–117.
Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making, 18*(1), 46.
Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant-DNA. *Scientometrics, 2*(4), 277–301.
Taşkın, Z., & Al, U. (2018). A content-based citation analysis study based on text categorization. *Scientometrics*, *114*(1), 335-357.
Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 103–110). Association for Computational Linguistics.
Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics, 12*(1), 393.
Valenzuela, M., Ha, V., & Etzioni, O. (2015, April). Identifying meaningful citations. In *AAAI Workshop*: Scholarly Big Data.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics, 7*(2), 272–285.
Xu, H., Martin, E., & Mahidadia, A. (2013). Using heterogeneous features for scientific citation classification. In *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*.
Zhang, P., & Koppaka, L. (2007, June). Semantics-based legal citation network. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law* (pp. 123–130). ACM.