Classifying the Ideational Impact of IS Review Articles: A Natural Language Processing Based Approach

Completed Research Paper

Julian Prester
School of Information Systems and Technology Management
UNSW Business School
UNSW Sydney
j.prester@unsw.edu.au

Gerit Wagner
University of Regensburg
Universitätsstr. 31, 93053 Regensburg, Germany
gerit.wagner@ur.de

Guido Schryen
Chair of Management Information Systems and Operations Research
Paderborn University, 33098 Paderborn, Germany
guido.schryen@upb.de

Abstract

By providing knowledge contributions and stimulating future research, review articles (RAs) play a vital role for cumulative knowledge development. Although many papers cite RAs, it is rarely transparent to which degree citation impact represents perfunctory citations as opposed to a deeper engagement with a RA's knowledge contributions. This distinction between perfunctory and ideational impact has largely been neglected in the literature arguably because of the manual effort required for qualitative analysis. Against this background, our study aims at developing automated classifiers of ideational impact of IS RAs. We propose a machine learning model based on natural language processing to evaluate the feasibility of automated analyses. The evaluation results provide evidence for an effective and scalable classification approach that presents a reliable and reproducible solution to the ideational impact classification problem. We discuss implications for improving the capabilities of understanding how IS scholars build on their field's body of knowledge.

Keywords: Ideational Impact Classification, Citation Content Analysis, Literature Reviews, Machine Learning, Natural Language Processing, Impact of Research Methods

Introduction

Review articles (RAs) play a vital role in the cumulative development of knowledge in the IS discipline. Since the prominent guidelines of Webster and Watson (2002), this has been reiterated in editorials (Rivard 2014; Rowe 2014), methodological guidelines (Paré et al. 2015; Schryen et al. 2015), and scientometric analyses (Wagner et al. 2016). In this discourse, the underlying assumption is that the value of RAs is rooted not only in their own knowledge contributions, such as synthesizing knowledge, building theory and aggregating empirical evidence (Schryen et al. 2015), but also in their capacity to stimulate subsequent research which, in turn, validates and extends their contributions. This integral function of facilitating cumulative knowledge development has been associated with the outstanding citation impact RAs have exerted on subsequent literature.

Even though many papers cite RAs to outline the general background and motivation of their research, it is rarely transparent to which degree citation impact represents perfunctory citations as opposed to a deeper engagement with a RA's knowledge contributions. For example, an analysis of the papers citing the review of Kohli and Devaraj (2003) showed that many citing articles (CAs) only briefly mention the review once in the background section. A few papers, however, report a deeper engagement with the actual contributions of Kohli and Devaraj. Most notably, Sabherwal and Jeyaraj (2015) provide a follow-up meta-analysis, published in *MIS Quarterly*, which extends the primary dataset, and validates and modifies the original theoretical propositions of Kohli and Devaraj. While the former examples neither challenge nor confirm the original RA, the latter example constitutes an exemplary case of cumulative knowledge development.

This significant difference has been emphasized repeatedly in the literature, which distinguishes so-called perfunctory from ideational impact (Hassan and Loebbecke 2017; Small 1978; Smith 1981; Takeda et al. 2011). Specifically, ideational impact refers to the uptake of a paper's ideas and concepts by subsequent research (Hassan and Loebbecke 2017, p. 18; Small 1978; Takeda et al. 2011, p. 3). Perfunctory impact, on the other hand, refers to citations that play only a minor role in the main argument of the CA (Hansen et al. 2006). Literature analyses have found that large fractions of citations are perfunctory (Bornmann and Daniel 2008; Hansen et al. 2006; Moravcsik and Murugesan 1975), and that this is a particularly common issue for highly cited works (Hassan and Loebbecke 2017, p. 10) such as review articles.

Although perfunctory impact generally does not contribute to the growth of knowledge in a discipline (Hassan and Loebbecke 2017, p. 10), the distinction between perfunctory and ideational impact has largely been neglected in the literature, in particular with regard to automated analysis. This leaves researchers with the task of applying qualitative methods of content, or citation context analysis, which requires considerable manual effort. Such a task is not only tedious and error-prone for humans but also difficult to accomplish reliably at a larger scale. In this context, analyzing the impact of RAs is an exemplary challenge, since RAs are among the most highly cited works in the IS discipline (Mingers and Xu 2010; Tahamtan et al. 2016). Thus whereas, manual citation content analysis might be feasible for papers that are not often cited, this approach becomes impractical for highly cited papers, research streams or even entire disciplines.

In this paper, we address the aforementioned issues by advancing state-of-the-art machine learning techniques to uncover the ideational impact of RAs. Specifically, we apply natural language processing (NLP) algorithms to the citation contexts of papers citing RAs. Although they are far from grasping all subtleties of natural language, these algorithms have been shown to be effective in various IS contexts (Abbasi et al. 2010; Larsen and Bong 2016). With these techniques, we pursue an analytical goal of classifying different types of impact. Put differently, these techniques allow us to assess the extent to which IS research cumulatively builds on RAs. Our approach thus provides a more adequate measure of cumulative knowledge development than traditional quantitative citation analyses, which do not distinguish between citations that reflect cumulative knowledge development and those that do not. We therefore address the following research question in this paper:

How effectively can NLP-based approaches classify the ideational impact of IS review articles?

As such, we offer three major contributions with this work. First, by adopting an NLP-based approach, we develop and evaluate a machine learning classifier and demonstrate how effective this automated analysis of ideational impact is. Second, following the importance of ideational impact, we suggest how NLP-based approaches offer opportunities for a qualitative analysis of this concept. Third, by highlighting the significance of ideational impact for the cumulative knowledge development in a discipline, we contribute to the discourse on different facets of impact that RAs exert.

Background

Ideational Impact and Cumulative Knowledge Development

The literature indicates that – beyond pure citation counts – different facets need to be considered when analyzing the citation impact of research output. A central assumption of many quantitative citation analyses is that citations indicate knowledge diffusion, knowledge growth, or cumulative knowledge development (Jackson and Rushton 1987; Wuchty et al. 2007). However, this view has been criticized

repeatedly. For instance, Smith (1981) emphasizes that a citation does not necessarily imply use and that not all citations are equal. Similarly, Bornmann and Daniel (2008) state that citations may have non-scientific reasons. In fact, Garfield and Merton (1979), the prominent creators of the Social Science Citation Index (SSCI), concede that citations "say nothing about the nature of the work, nothing about the reason for its utility or impact" (p. 246). These differences in citations have led several methodologists to call for qualitative methods that take into account these significant differences between citations (MacRoberts and MacRoberts 1989; Nicolaisen 2007). Citation types and corresponding reasons to cite have been classified in several ways (Bornmann and Daniel 2008; Hansen et al. 2006; Moravcsik and Murugesan 1975). A major challenge in empirical studies is that their conceptions of impact types vary considerably, inhibiting the replicability of their results (Bornmann and Daniel 2008). Furthermore, there are very few automated approaches to classify citation types (Lu et al. 2017).

The distinction between perfunctory and ideational impact can be made explicit by differentiating citations that do not contribute to the 'growth of knowledge in a discipline' (Hassan and Loebbecke 2017, p. 10) from those that cumulatively add to an existing body of knowledge. Specifically, Hassan and Loebbecke (2017) contrast perfunctory impact with ideational impact. While the former refers to citations that do not play a significant role in the main argument of the citing paper (Hansen et al. 2006), the latter refers to the uptake of ideas and concepts by subsequent research (Hassan and Loebbecke 2017, p. 18). In this context, different scientometric studies have found varying yet, consistently large proportions of perfunctory citations (Bornmann and Daniel 2008; Lu et al. 2017; Serenko and Dumay 2015). They have also highlighted the effort required for manual content analyses, the challenges of coding citation types reliably, and the problem of handling erroneous references caused by citation indices. While this distinction between perfunctory and ideational impact has been suggested to be critical for analyzing cumulative knowledge development, we are not aware of any machine learning approach that addresses the above-mentioned challenges by distinguishing ideational and perfunctory impact.

Natural Language Processing and Text Classification

Advances in the availability of digital text and computing processing power have opened up new opportunities for automated analyses of natural language both in written and spoken form. The application of NLP and text classification has flourished recently (Jurafsky and Martin 2009), and both methods are suitable for assessing links between documents, as represented through differences in the use of language. Thus, tasks and challenges presented in the previous subsection, such as content analyses of scientific documents, reliable coding of citation types, and overcoming subjectivity and variation in qualitative impact analyses, are considered text classification problems, which NLP-based approaches can help to solve (Dong and Schäfer 2011; Stremersch et al. 2015; Teufel et al. 2009).

Given the complexity of human language, NLP approaches typically split language into seven conceptual levels (Mitkov 2005): phonology, morphology, lexicography, syntax, semantics, discourse, and pragmatics. The syntactic and semantic levels are of particular importance for text classification. NLP algorithms break text into multiple components, which can then be used to model the way in which the semantic meanings of phrases and sentences are derived from the meanings of their syntactic constituents. On the syntactic level, NLP algorithms examine how words are combined and used to form sentences. Major tasks on the syntactic level include word segmentation, stemming, part-of-speech (POS) tagging, and parsing (Mitkov 2005). To identify the meaning of what is written, the semantic level goes beyond the structure of words and sentences. Semantic analysis is considered as a first step toward natural language understanding (Mitkov 2005). Prominent applications on the semantic level include machine translation, sentiment analysis, named entity recognition, and topic recognition.

Text classification has always been one of the major tasks of qualitative text analysis (Holsti 1969; Neuendorf 2002). Text classification is defined as the clustering of similar texts or documents into discrete categories. Such a classification is typically based on a large number of features characterizing the text (Martens and Provost 2014). While text classification tasks have traditionally been conducted manually, today, modern machine learning algorithms can automate much of the manual effort involved in coding text and placing it into categories. Thus, this technique presents a reliable alternative to manual qualitative document analyses and as such is the prevailing methodological approach for citation classification (Dong and Schäfer 2011; Stremersch et al. 2015; Teufel et al. 2009).

Proposed Methodology for Ideational Impact Classification

We approached our research question based on three consecutive steps. First, we gathered and coded citation sentences from CAs and their corresponding RAs (see Figure 1). Second, based on these citation sentences and the cited papers (i.e., the RAs), we employed NLP-based techniques to derive a set of syntactic, semantic, and contextual features representing certain citation patterns (see Table 2). Finally, we used the ideational impact coding and the NLP-based feature set to develop three classification models (see Figure 3). Next, we provide a detailed explanation of the three steps.

Data Collection and Coding Procedure

In order to develop and evaluate the ideational impact classification method, we collect a document corpus comprising RAs on IT business value and their CAs. Our full sample of RAs is based on the set described by Wagner et al. (2016). They originally identified 214 standalone RAs that have been published in a set of 40 major IS journals between 2000 and 2014. From this dataset, we consider RAs in the domain of IS business value, which leaves us with a total of 22 RAs. We selected this context with the reliability and generalizability of our classification method in mind. More specifically, to ensure adequate reliability when manually coding the document corpus, we selected a familiar research domain for which we were able to code the ideational impact of citing papers consistently. The domain is mature enough (Dehning et al. 2004; Schryen 2013) to provide sufficient number and diversity of RAs (e.g., theoretical RAs and meta-analyses) as well as a sufficient number of CAs. Although evaluations in further domains are necessary, we are confident that this mature domain provides a suitable context for our study.

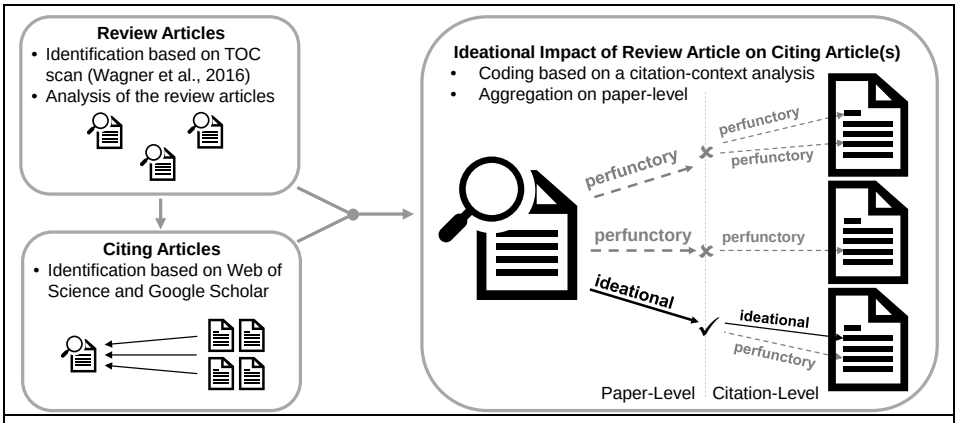

Figure 1. Data Collection and Coding Procedure

Based on this literature sample, our methodology combines multiple steps of citation and content analyses, as shown in Figure 1. We conducted a citation analysis in the form of a forward search to find all CAs potentially using the knowledge developed in the cited RA(s). We gathered data from two literature databases that provide detailed citation data, i.e., Google Scholar and Web of Science. Google Scholar, lauded for its comprehensive coverage, has been used extensively as a citation data source in previous studies (Bornmann and Daniel 2008; Grover et al. 2013; Judge et al. 2007). To improve data quality, we ran the same forward search for each RA on Web of Science and merged the two exports. As the merged export contained some inconsistencies, a removal of duplicate entries and imputation of incomplete data was, as far as possible, necessary. The steps resulted in the identification of approximately 30,000 CAs. Since the generation of an annotated corpus for the training of the machine learning classifiers requires a manual coding of every paper, the full dataset was too large to be coded manually. Hence, we decided to filter the CAs for papers that have been published in journals included in the Senior Scholars' Basket of Journals. This set of eight journals is widely acknowledged as a basket of top journals in the IS field and recognizes topical, methodological, and geographical diversity. Thus, the final dataset comprises 1,228 CAs published in journals included in the Senior Scholars' Basket of Journals in addition to 22 RAs on IS business value.

Based on the distinction between perfunctory and ideational impact, we employ our coding scheme in line with Hassan and Loebbecke's (2017) understanding of citations that cumulatively add to an existing body

of knowledge. Relationships that we considered when coding ideational and perfunctory impact are summarized in Figure 2. The ideational impact was coded by manually analyzing each of the 1,228 CAs. We coded ideational impact when we found an explicit and direct attribution of the impact a RA has had on the CA within the citation context (Valenzuela et al. 2015). Thus, when coding ideational impact it is necessary to consider the text of both the citing and cited documents to be able to make a judgment as to whether a citation does indeed represent use of material covered in the cited document (Smith 1981).

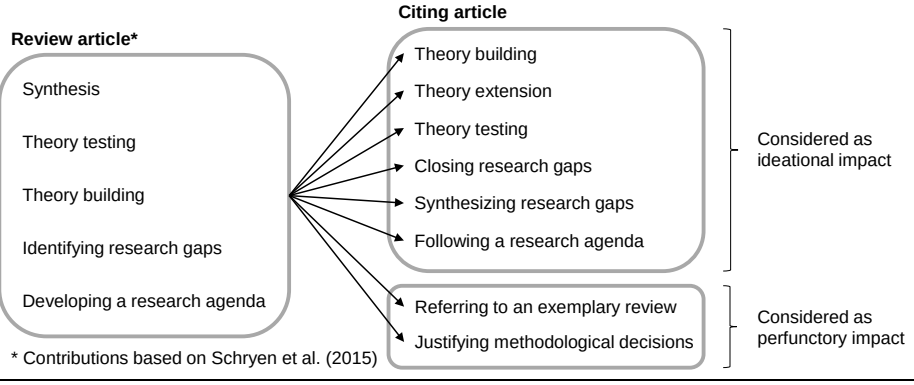

Figure 2. Relationships Considered Between RA and CAs

While ideational impact was coded on a paper-level, the machine learning features were developed on a citation-level and aggregated to the paper-level, representing a nested sample of paper-level coding and features. Table 1 provides examples from our coding, illustrating both impact types.

Table 1. Exemplary Citation Sentences by Impact Type

Impact Type	Representative Citation Sentence (in the Citing Article)	Rationale for Coding
Ideational	"Drawing on the resource-based view (RBV) of the firm as an overarching framework and prior research ([...] Melville et al. 2004 [...]), we propose three reasons to explain why overall IT investments are likely to have a positive association with accounting profits." – (Mithas et al. 2012, p. 207)	CA explicitly states that it draws on the RBV making it a case of theory building that aligns with existing frameworks.
	"This decision was based, in part, on work suggesting that our understanding of the BVIT would benefit from the use of primary data to empirically examine the link between IT and firm performance ([...] Wade and Hulland 2004)." – (Nevo and Wade 2011, p. 408)	CA explicitly takes up the suggestions made in the RA making it a case of following a research agenda.
Perfunctory	"DeLone and McLean (1992) provide a thorough overview of the main research in the quest for the key success factors of that time." – (Bartis and Mitev 2008, p. 113)	RA is cited as an exemplary review on the topic.
	"Information technology (IT) that promises to enhance organizational performance costs companies millions of dollars to implement (Kohli and Devaraj 2003)." – (Xue et al. 2011, p. 400)	RA is cited to highlight the business impact of the general topic.

To ensure reliable results between the two coding authors, we implemented a three-phase coding process consisting of (1) a training phase, (2) a reliability assessment phase, and (3) an individual coding phase (Neuendorf 2002). We first coded an initial subsample to develop an understanding of the coding task and to refine coding guidelines. We then coded a larger sample to compare the results of the individual coders and evaluate our coding in terms of reliability. Discrepancies between the two main coders were reconciled by a third coder. Finally, two authors individually coded the remaining dataset. The two individual coding sets as well as the reconciled reliability coding and initial training coding were then used as the "true" coding of ideational impact for the automated classification task and as such comprise the

full sample. Inter-rater reliability was sufficient with a Cohen's Kappa value of 0.89 for the ideational impact coding of the reliability sample. The final reconciled set resulted in a distribution of 30% cases of ideational impact and 70% cases of perfunctory impact. Note that the coding procedure was completed at an early stage of the research project before the construction of the feature set and therefore cannot have affected the development of our machine learning models or corresponding feature sets.

Feature Set Construction

For a machine learning approach to be able to identify ideational impact, a set of features is necessary for the classifier to determine both the citation sentences' underlying characteristics and the characteristics of the cited RA. As such, when we consider our ideational impact coding as the outcome variable, the feature set comprises variables that predict ideational impact. As shown in Table 2, we develop our feature set, based on the syntactic and semantic dimensions of natural language. Additionally, we consider contextual features comprising metadata of the CAs. To improve the reliability of our dataset and the robustness of our analysis, we performed various pre-processing steps, before starting the data extraction from the citation sentences and their contexts. In line with the recommendations of Debortoli et al. (2016) our pre-processing comprised the following steps: collection of the full-text documents of all CAs, filtering of impractical documents, citation sentence and citation context (preceding and subsequent sentences) extraction, and word segmentation. Since ideational impact depends not only on the CA, but also on the cited RA, we construct features from both documents.

Syntactic Features: Since we expect citation sentences that signal ideational impact to adhere to a specific sentence structure we operationalize a set of syntactic features. By analyzing the *textual type* of a citation, we distinguish between the two major syntactic types, namely textual (author name outside the reference marker) and non-textual (simple reference marker) citations (Valenzuela et al. 2015). We further look at whether the citation '*stands alone*' or whether it is a part of multiple references grouped together in one reference marker. Additionally, as proposed by Jochim and Schütze (2012), we extract the absolute *position of the citation* within the citation sentence ranging from 0 (first word of the sentence) to 1 (last word of the sentence). We also developed features based on the POS sequences of the citation sentences. We derive features representing the use of *comparative and superlative clauses* (e.g., more, best), first and third person *personal pronouns*, and *POS patterns* signaling certain citation functions, as described by Dong and Schäfer (2011). The POS patterns are formulated as regular expressions that match certain syntactic structures of a citation sentence.

Semantic Features: Topic relatedness between citing and cited papers, it is argued, is an important indicator of ideational impact. In line with Guo et al. (2014), we operationalize topic relatedness between citing and cited papers as *title and abstract similarity*. We develop two LSA models (one for document titles and one for document abstracts) to derive topic relatedness values between RAs and CAs (Larsen et al. 2008; Sidorova et al. 2008). To construct comprehensive and coherent topic models, we followed the guidelines of Debortoli et al. (2016) in using different dimensionality reduction techniques to evaluate the best performing LSA models for our citation dataset. The data preparation steps included: collection of unigram, bigram, and trigram word segmentations (e.g., "IS", "business", and "value" or "IS business" and "business value" or "IS business value"), removal of punctuation marks and other technical symbols (e.g., periods, commas, "@", or "©"), removal of numbers (e.g., "2004"), removal of common stop words (e.g., "the", "of", or "and"), stemming (e.g., reducing "study", "studies", and "studying" to their base form "study"), and filtering for terms fulfilling a certain POS function (e.g., removing nouns, verbs, or adjectives).

Another semantic feature that has been proven useful for citation classification is document *sentiment* (Athar 2011; Jochim and Schütze 2012). Sentiment analysis tries to extract affective states and subjective information from documents to determine the author's attitude toward the text. Hence, we include measures for citation sentence and citation context sentiment to control for author's attitude toward the RA, which is expected to be more extreme in ideational impact cases (e.g., either very positive or very critical) (Athar 2011; Jochim and Schütze 2012). We derive negative, neutral, positive, and aggregated sentiment scores for each citation sentence and citation context.

Lastly, we consider the manually coded *knowledge contributions* of the RAs as a prerequisite for ideational impact on the CAs. In essence, we consider different types of knowledge contributions as identified by Schryen et al. (2015). In their content analysis of RAs, they identify the following types of

knowledge contributions: synthesis, adoption of a new perspective, theory building, theory testing, identification of research gaps, and provision of a research agenda. This consideration of the RA in the feature set is important since closing of research gaps, following a research agenda, or theory testing is only possible for CAs if the corresponding knowledge was developed by the cited RA (Schryen et al. 2015; Smith 1981).

Table 2. Feature Set

Feature	Description (Source)	References
Syntactic Features		
Textual type	Number of textual citations (CA)	Valenzuela et al. (2015)
'Standalone' reference	Number of 'standalone' citations (CA)	-
Position in sentence/context	Position of the reference in citation sentence/context (CA)	Jochim and Schütze (2012)
Comparative/superlative clauses	Number of comparative and superlative clauses (CA)	Jochim and Schütze (2012)
Personal pronouns	Number of personal pronouns (CA)	Jochim and Schütze (2012)
POS patterns	Appearances of POS patterns (CA)	Dong and Schäfer (2011)
Semantic Features		
Title/abstract similarity	Semantic similarity of the titles and abstracts (RA & CA)	Guo et al. (2014)
Sentiment (positive, negative, neutral, aggregated)	Sentiment of the citation sentence and context with regards to the RA (CA)	Athar (2011), Jochim and Schütze (2012)
RA knowledge contributions	Knowledge developed in the cited RA as a prerequisite for ideational impact (RA)	Schryen et al. (2015)
Contextual Features		
Citations in introduction/ background/main sections	Number of citations in the introduction, background, and main section (CA)	Dong and Schäfer (2011), Jochim and Schütze (2012)
Citations toward the RA	Total number of RA citations (CA)	Jochim and Schütze (2012)
Citation sentence/context variety	Number of different citations in the citation sentence/context (CA)	Dong and Schäfer (2011), Jochim and Schütze (2012)
Citation sentence/context density	Focal citations divided by total citations in the citation sentence/context (CA)	Dong and Schäfer (2011), Jochim and Schütze (2012)
Total number of references	Total number of references in the CA's bibliography section (CA)	Jochim and Schütze (2012)
Total citations	Total number of citations in the CA (CA)	Jochim and Schütze (2012)
Weighted citation count	RA citations divided by total citations (CA)	Jochim and Schütze (2012)
Self-citation	At least one author of the RA and CA is identical (RA & CA)	Wan and Liu (2014), Teufel et al. (2009)

Contextual Features: The development of our contextual features is mostly based on citation metadata from the RAs and their CAs. Since the location of a citation has been shown to be "the most reliable information on citation function one could obtain from the paper directly" (Dong and Schäfer 2011, p. 625), we extract both the *location* of the citations in the CA (i.e., introduction, background, or main section) as well as the total number of *citations toward the cited RA*. We expect that citations appearing in the main sections (i.e., all sections following the background section such as method, results or discussion sections) of a CA are more likely to signal ideational impact than those appearing in the introduction or background section. We further extract the number of different references cited in the citation sentence and context (i.e., *citation sentence/context variety*) and the proportion of citations

toward the RA against total citations in the citation sentence and context (i.e., *citation sentence/context density*). We also include the *total number of references* in the bibliography section and the *total number of citations* of all cited papers. This is done to derive a *weighted citation count* over all references. Since citing one's own research might indicate a higher probability of re-use of intellectual material from previous work, we include a binary feature indicating a *self-citation* (Wan and Liu 2014).

Ideational Impact Classification

When developing our classification models, we focused on effectiveness of the NLP cased classifiers as our primary evaluation criterion due to the exploratory nature of our study. We also explored other characteristics of the classifiers such as interpretability. However, these were not the primary focus when developing the model. We discuss potential extensions with a focus on explanatory insights into the effectiveness and interpretability of individual features in the future research section.

We implemented a range of state-of-the-art classification algorithms such as logistic regression, naïve bayes, support vector machine, random forest, gradient boosting, and eventually an ensemble method combining three of the individual classifiers. We focused our selection of classification methods on algorithms that are capable of performing binary classification tasks and that have been shown to be successful in earlier research on citation and text classification (Abbasi et al. 2010; Dong and Schäfer 2011; Guo et al. 2014; Martens and Provost 2014). We then evaluated and compared the classification algorithms against a baseline classifier, against each other, as well as against human experts. In the following, we present the results of three binary classification models to distinguish ideational impact: a baseline classifier, a logistic regression classifier, and an ensemble machine learning classifier. All three models have in common that they use ideational impact coding as the binary outcome variable. The three models are described in detail in the following.

A baseline classification was implemented (see Figure 3a) prior to using machine learning algorithms to develop a model. Such a baseline model (*model 0*) is useful to define a minimal classification performance, which is necessary to gauge the performance of more sophisticated models. The baseline classifier is based on a simple guessing of the most probable impact type, which, in our case, is perfunctory impact (70% of cases). Thus, the model is only using the ideational impact coding and the underlying impact distribution as an input.

Logistic regression (*model 1*) was then used to provide a first attempt to make use of the NLP-based feature set (see Figure 3b). As such, logistic regression is used to classify the probability of a citation representing ideational impact. In addition to the ideational impact coding, which serves as the dependent variable, model 1 includes the knowledge contribution features from the RAs as well as the syntactic, semantic, and contextual features from the CAs as independent variables. To prevent the logistic regression classifier from overfitting and failing to classify reliably hitherto unseen RAs and CAs, we further introduced a split of the full dataset into a training and a test set. The logistic regression model was first fitted using the training set and then evaluated against the test set. To deal with the highly imbalanced distribution of the outcome variable in our dataset, we use a stratified sampling approach to ensure that relative class frequencies are preserved when splitting the dataset.

The third and most complex model (*model 2*) is an ensemble machine learning classifier (see Figure 3c). Generally, the benefit of ensemble methods is to combine the predictions of several individual classifiers, which are built with base classifiers to improve generalizability and robustness over a single classifier. Therefore, we combine conceptually different classifiers and use the average predicted probabilities of the individual classification results to predict the class outcomes. Such an ensemble classifier based on a weighted voting can be useful for a set of similarly well-performing classifiers to balance out their individual weaknesses. As such, model 2 utilizes three algorithms, which have all been evaluated individually on the classification task, to aggregate classifications based on an optimized weighting scheme. Specifically, the ensemble consists of three state-of-the-art machine learning classifiers: a support vector machine, a random forest classifier, and a gradient boosting classifier. The individual classification results enter the final outcome based on the optimized weighting scheme of 11%, 66%, and 22% respectively. The optimal weights were identified through a grid search over the potential weighting scheme candidates. A pilot test assessing the performance of each individual machine learning classifier showed promising results. However, a paired permutation test showed that the optimized ensemble classifier performed significantly better ($p \leq 0.05$). Including classifiers other than the three mentioned

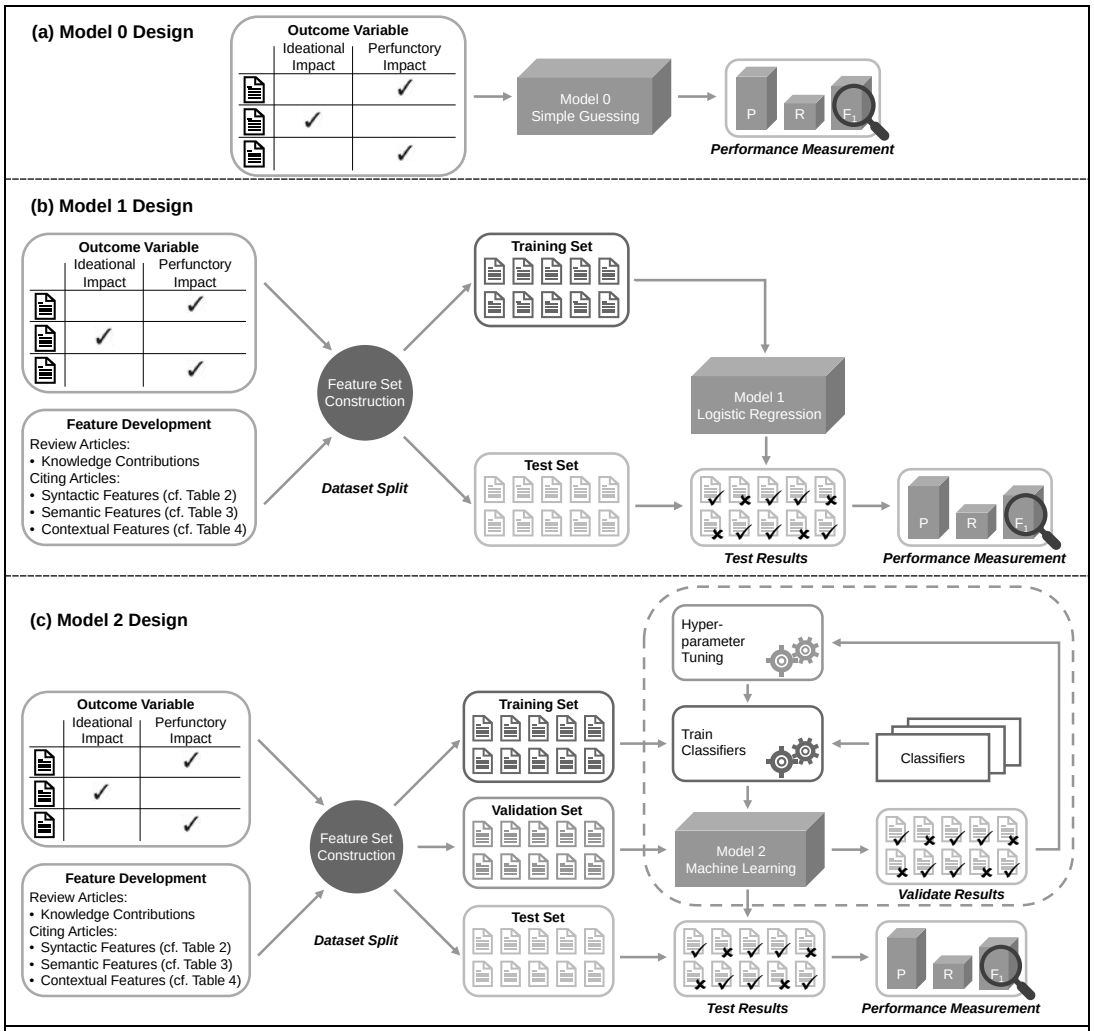

Figure 3. Ideational Impact Classification Models

into the ensemble did not show any evidence of further improving classification effectiveness. The model comprises three classifiers with a weighted impact on the overall classification result, the ideational impact coding as the outcome variable, and the knowledge contributions, syntactic, semantic, and contextual features as the feature set for the machine learning classifiers. Since the machine learning classifier uses parameters to tune classification outcomes, we applied iterative model fitting and testing procedures to identify the optimal set of parameters for the model. The model optimization necessitated not only a division of the dataset into a training and test set (as in the case of model 1) but also a split into a third validation set to evaluate each parameter setting (Han et al. 2011). We used 10-fold cross-validation to evaluate the performance of the fitted models.

Evaluation and Findings

We first present our findings with regards to classification effectiveness of the three models as evaluated by established measures. We then show how the results change for different contexts depending on the objective of the classification. Finally, we briefly go beyond our main objective of model effectiveness and discuss the interpretability of our model by analyzing the relative importance of the different features for the classification task.

We trained each of the three models on the full dataset with ideational impact as the binary outcome variable. The evaluation results for the 80-20 train-test-split are given in Table 4. Table 4 also contains the annotation performance of human experts for a comparison. We derive a score representing human

expert annotation using the disagreement from the reliability coding against the reconciled coding, which we posit as the "true" coding of ideational impact. While we acknowledge that future research should evaluate performance against a third-party human expert not involved in the coding process, prior research has discussed evaluation against human coders as an initial benchmark. For example, Jochim and Schütze (2012), Pappas and Popescu-Belis (2016), and Pham and Hoffmann (2003) have all provided evidence for the effectiveness of using human coding results as a benchmark. As such, we take the reconciled coding as the "true" coding of a RA's ideational impact on a CA and derive a score for human expert annotation performance by measuring the discrepancy between this "true" coding and the divergent human coding before reconciliation. Since the highly imbalanced dataset renders performance measures such as classification accuracy ineffective, we report the outcome of each of the three models using precision, recall, and F_1-scores (see Table 4) (Han et al. 2011). These three performance measures are derived from confusion matrices for the ideational impact classification task. A sample of the model 2 classification results is presented in Table 3. The matrix shows that when evaluated against the test sample of 152 cases, model 2 predicts that 25 are ideational impact cases (of the 30 actual ideational impact cases), and that 101 are perfunctory impact cases (of the 122 actual perfunctory impact cases).

Table 3. Ideational Impact Confusion Matrix (Based on Results of Model 2)

	Actual Ideational Impact	Actual Perfunctory Impact
Predicted Ideational Impact	25 (TP)	21 (FN)
Predicted Perfunctory Impact	5 (FP)	101 (TN)

In the context of ideational impact classification, *precision* describes the number of correctly classified ideational impact cases (true positives (TP)) among all classified ideational impact cases (TP + false positives (FP)). Thus, *recall* describes the number of correctly classified ideational impact cases (TP) among all true ideational impact cases (TP + false negatives (FN)). The *F_1-score* is then defined as the harmonic mean of precision and recall, which offers an integrated measure to balance the trade-off between precision and recall. Further, we present receiver operating characteristic (ROC) together with their corresponding area under the curve (AUC) (Sinha and May 2004). ROC curves provide complementary insights as they not only measure classification performance but also visualize. The visualization is achieved by plotting a binary classifier's true positive rate against its false positive rate at various discrimination threshold settings (Fawcett 2006).

Table 4. Performance of Classification Models Against Human Expert Annotation

Model	Precision	Recall	F_1-score
Model 0 – Simple Guessing	0.30	1.00	0.46
Model 1 – Logistic Regression	0.71	0.48	0.57
Model 2 – Machine Learning Classifier	0.83	0.54	0.66
Human Expert Annotation	0.52	0.87	0.65

Table 4 shows the results for the three models compared to human expert annotation. The best results as measured by the F_1-score were obtained using the machine learning classifier (model 2). This model also provides the best balance between precision and recall. Comparing model 2 with human annotation further confirms that the machine learning model performs on par with human experts. However, it is interesting to note that while precision is high for model 2 and low for human expert annotators, the recall measure shows opposite results. The logistic regression classifier (model 1) has a lower overall F_1-score, as well as lower precision and recall scores. As such, model 1 performs worse than both human experts and model 2 when measured by the F_1-score. Precision and recall are again inversely balanced. Although the baseline classifier (model 0) achieves a predictive accuracy of 71%, the F_1-score illustrates that this is only because we deal with a highly imbalanced outcome variable (30% ideational impact and 70% perfunctory impact, as outlined previously). In this manner, precision and recall are also highly imbalanced in the case of model 0. Thus, when measured by F1-scores, model 0 is outperformed by both model 1 and model 2.

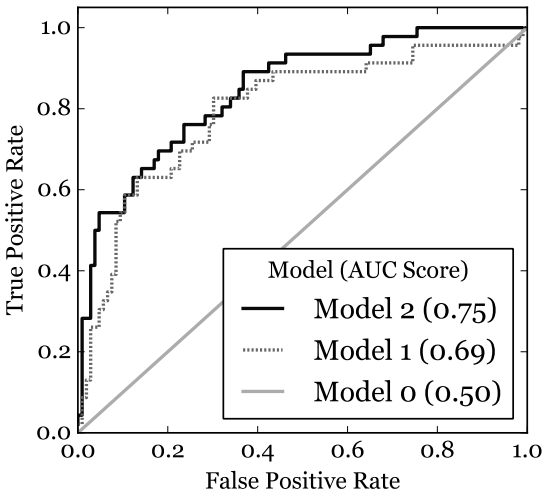

Figure 4. ROC Curves of the Ideational Impact Classification Task

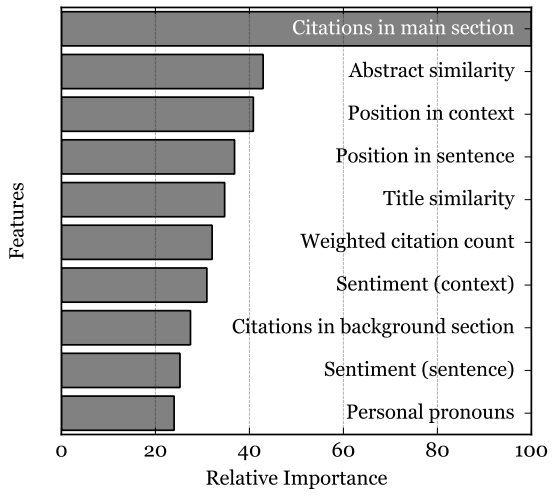

Figure 5. Relative Feature Importance Measures

Figure 4 shows the ROC curves, illustrating the predictive performance of our three models. Curves situated closer to the top left corner signify better results, since they denote high ratios of true to false positives. Looking at the figure, we can see that model 2 has the best classification performance, followed by model 1. The AUC for the model 2 ROC curve amounts to 0.75, which means that the model can distinguish between a randomly drawn ideational impact citation and a randomly drawn perfunctory impact classification with an accuracy of 75% (Fawcett 2006).

In summary, the results along the three performance measures precision, recall, and F_1-score lead us to engage in a more differentiated discussion of our ideational impact classification approach. In general, while the precision results for the three models are better than those for the human annotators, the human annotators outperform the three models in terms of recall. These findings offer implications concerning the context in which the ideational impact classification method can be used.

While the above evaluations confirmed the effectiveness of the ideational impact classification method, it can be difficult to interpret the output of machine learning classifiers beyond their predictive accuracy. Thus, next we evaluate the overall best-performing model in more detail and assess the most effective features contributing to its performance. We implement this analysis by examining the relative importance measures of the features for the classification outcome. Since the ensemble machine learning classifier is based on random forest and gradient boosting estimators, we can use the relative rank of a feature used as a decision node to assess the importance of that feature for the predictability of the target variable. Features appearing at the top of the tree are generally attributed higher weights for the classification decision (Breiman et al. 1984). Specifically, the relative importance measures indicate how influential each feature is compared to the most important feature for the classifier. We evaluated relative feature importance and robustness of the measures derived from the random forest and gradient boosting classifiers and compared the results between the two and also against the effects of the variables from the logistic regression classifier as measured by regression coefficients. The feature importance scores showed some volatility due to the different regularization mechanisms (the sampling of features every time a new decision tree is grown) between the two classifiers. However, the most important features were largely consistent regardless of the machine learning classifier. Similarly, when comparing the feature importance measure with the logistic regression coefficients we got almost identical results with eight out of the top ten features appearing as the most influential and significant ($p \leq 0.05$) coefficients.

Figure 5 shows the ten most influential features for classifying the ideational impact of a citation as derived from the gradient boosting classifier. The chart shows that there is one feature that dominates in terms of how effective it is for the classifier to make predictions. The "Citations in main section" feature, which describes the number of citations appearing in the main sections of the paper, is more than twice as effective in the classification of ideational impact than the second most important feature. For example,

the feature "Abstract similarity" can be interpreted as being 43% as influential for the classification model as the most important feature. Further, the contextual feature "Citations in background section", which is related to the group of features covering "Citations in introduction/background/main sections", can also be found among the Top-10. In addition, the two LSA model features (i.e., "Title similarity" and "Abstract similarity") strongly impact classification performance. Another noteworthy finding in terms of feature importance is that none of the genre-specific human-coded features appear among the high-performing features as measured with either of the three mentioned methods.

Discussion

Classifying and Assessing Ideational Impact

Our machine learning based classification method allows us to assess the feasibility and effectiveness of classifying the ideational impact of RAs. While recall rates could be improved in future research, the precision and overall classification performance are convincing. This enables the reliable classification of ideational impact, thereby offering new methodological opportunities for future literature analyses. Due to the exploratory nature of our study, we discuss our findings in a way to progress toward more formalized knowledge by formulating three propositions.

Regarding classification effectiveness, our results show that the proposed method provides reliable, conservative classification, which can distinguish between ideational impact and perfunctory impact citations with an accuracy of 75% (see AUC measures in Figure 4). However, there is still potential to improve recall rates. To illustrate this, we distinguish three classification settings. First, if users are interested in a conservative classification of ideational impact, they prefer a low false positive rate and are tolerant toward false negatives. In this setting, in which high precision rates are preferred, the proposed classification method is generally more effective than human annotators. Second, if users are interested in identifying all instances of ideational impact instead, they prefer a low false negative rate and are tolerant toward false positives. This is the case when missing ideational impact must be avoided at the cost of manually analyzing papers that may turn out to be false positives. In this setting, in which high recall rates are preferred, human annotators outperform the proposed classification method. Third, if users are interested in a balanced classification outcome, they favor neither recall nor precision (and the respective false-positive and false-negative rates). In such settings, in which the focus is on the overall classification performance, the human annotator and our proposed machine learning classifier perform on par. Thus, we present the first proposition:

P1: NLP-based machine learning classifiers are a viable option for distinguishing an ideational impact citation from a perfunctory impact citation when an accuracy of 75% is sufficient.

The results of the machine learning classifier suggest that such designs can be effective in classifying ideational and perfunctory impact. To achieve this classification performance, the proposed classifiers draw on a broad set of syntactic, semantic, and contextual features that are constructed both from the citing and the cited document. The capability of our machine learning algorithms to classify ideational and perfunctory impact reliably and conservatively constitutes an important contribution to the methodological arsenal of papers analyzing citation types. These analyses have heretofore suffered from low replicability with percentages of perfunctory citations ranging between 10% and 50%, depending on how the authors define perfunctory impact (Bornmann and Daniel 2008). Effective classification techniques do not only have the potential to affect a major shift toward more replicable research results in this literature; they can also be applied on a larger scale to uncover which citations, and thus citing papers, contribute to the growth of knowledge (Hassan and Loebbecke 2017). These aspects of reliability and replicability become particularly important considering that the genre-specific human-coded features, which required a considerable human coding effort, did not show any evidence of being among the highest-scoring features in terms of feature importance. This finding seems to be at odds with the prominent argument in the literature on citation behavior that to identify ideational impact one has to consider both the content of the citing as well as the cited article (Smith 1981). We would not, however, consider our results to be a conclusive rejection of this argument since other features from the focal paper may still afford significant improvements in classification performance. Generally, our insights may be useful to guide citing decisions toward literature reviews that have exerted a higher ideational impact on

the literature. Furthermore, aggregated figures of ideational impact may be used to assess and compare the knowledge-related impact of different RAs. Therefore, the second proposition becomes:

P2: NLP-based approaches are a reliable and replicable alternative to manual citation analyses.

On an aggregated level, we observe that ideational impact differs considerably from citation counts. The importance of distinguishing ideational impact from overall citation counts has been emphasized repeatedly by methodologists (Moravcsik and Murugesan 1975), who have – to the best of our knowledge – not provided quantitative evidence that substantiates their critique. As a result, this critique has often been dismissed in IS papers, which assume that while not every citation may reflect ideational impact, aggregated citation counts are still a good indicator for ideational impact, or knowledge development (Cuellar et al. 2016). Our dataset shows that there are only moderate to week correlations between aggregated ideational impact and citation counts when some of the highly cited RAs are excluded as outliers. Further research that addresses the limitations of pure citation counts by distinguishing ideational from perfunctory impact is therefore warranted. Hence, the third proposition states:

P3: On an aggregated level, ideational impact measures provide a better understanding of cumulative knowledge development than citation counts.

With these insights in mind, our paper contributes to the broader discourse on the impact of academic literature (Hassan and Loebbecke 2017), cumulative knowledge development (Keen 1980), and research assessment (Truex et al. 2009) in IS. We think that our proposed approach offers a solution to an important and relevant problem for research practice not just in IS but in most social science disciplines. Citation classification is the basis for large parts of our research evaluation spanning levels of papers, authors, and journals all of which can benefit from our proposed methodology. Further, citations play an increasingly important role for our search tools in everyday research. As Hassan and Loebbecke (2017) argue, citations are not only the most efficient way of searching for relevant studies across disciplines, but also to effectively position the relevance of IS amongst other disciplines. Although various techniques to evaluate the impact of a paper exist, the most frequently used method is citation count analysis (Bornmann and Daniel 2008). While citation indices such as Google Scholar and Web of Science made the traditional approach of citation analysis deceptively simple, our study demonstrates that distinguishing perfunctory from ideational impact is necessary to assess the impact of papers more appropriately. Furthermore, our proposed method based on NLP could be beneficial beyond the IS discipline and be applied in other business disciplines and the social sciences more broadly. While we do not claim that citations serving perfunctory purposes are not important for writing a research manuscript, we agree with Hassan and Loebbecke (2017) that perfunctory citations do not contribute to the cumulative knowledge development of a discipline, especially when compared to citations indicating ideational impact. By developing an ensemble of machine learning classifiers that is effective in classifying ideational impact, we improve the discipline's capabilities of understanding how IS scholars build on their field's body of knowledge to make sure we are "correctly perched on the shoulders of the giants" (Serenko and Dumay 2015, p. 1350).

Future Classifier Improvements and Application Areas

We see areas for future research as two consecutive stages of firstly improving the performance of the classification model in general and secondly applying the classifiers to address research questions in specific application areas. We present this proposed path to future research in Table 5 and discuss both stages in detail in the following paragraphs.

In the first stage we call for research to improve and extend the ideational impact classifiers. We provide evidence that our NLP-based approach is indeed capable of distinguishing an ideational impact citation from a perfunctory impact citation making it a reliable and replicable alternative to manual citation analyses. Nevertheless, our findings also reveal areas for improvements in terms of performance and scope. Therefore, we call for future research to develop further our understanding of citation behavior and to further define operational criteria for distinguishing ideational and perfunctory impact. A more detailed and codified understanding of the key concepts will eventually lead to more robust classifiers and improved classification performances. We believe that our approach can be transferred to other disciplines beyond IS. By limiting the scope of our study in terms of the selected time frame (papers published between 1993 and 2014), genre (IS literature reviews), and publication outlets (40 major IS

journals), we call for future studies to analyze how well our classifiers perform on different datasets. Although our sample is a major dataset regarding the number of papers analyzed, the sample is still relatively small for the training of machine learning classifiers. Future research may evaluate other machine learning classifiers in a broader context, considering different research genres, domains, and disciplines. Further, the development of new and fine-tuning of established features can help to develop more robust models and increase classification performance. Promising paths to develop new features include, for example, semantic features based on the full text of academic literature, and further features representing the relationship between RA and CA. Even unsupervised learning and topic modelling techniques, such as latent dirichlet allocation (Larsen et al. 2008; Sidorova et al. 2008), could be applied in an exploratory study to inductively develop citation types and features grounded in the citation data.

Table 5. Future Classifier Improvements and Application Areas

Stage 1	Classifier Improvement	Research Questions
	Performance	How can the concept of ideational impact be refined to enhance classifiers? How can new and refined NLP features improve classification performance? Are unsupervised classification models better suited to classify ideational impact?
	Scope	How can the classifier be applied and extended to classify ideational and perfunctory impact for different genres, domains, and disciplines?
Stage 2	Application Area	Research Questions
	Research evaluation	How can rankings based on ideational impact help to align research evaluation with actual knowledge development by excluding perfunctory citations?
	Scientometric analyses	How can ideational impact serve as new dependent variable in scientometric impact models? How do science maps such as those resulting from co-citation analyses change when the influence of perfunctory impact is removed?
	Search tools based on citation data	How can ideational impact measures be included in existing approaches to literature search? How can the effectiveness of literature search tools be improved by including ideational impact measures?

In the second stage we call for scholars and practitioners to put into use the improved ideational impact classification models in applications that are being utilized in academic practice. Three immediate application areas are research rankings, scientometric analyses, and search tools based on citation data.

Today, rankings and evaluation of papers, authors, journals, institutions, and entire disciplines are largely based on citation data. Research evaluation in most institutions relies on citation indices (e.g., the Journal Impact Factor), which are produced by industry-scale companies (e.g., Clarivate Analytics). Distinguishing ideational from perfunctory impact could foster the development of new measures of citation impact for papers, authors, and journals. These measures could support different stakeholders in evaluating research output in a way that is less susceptive to well-known weaknesses of citation analysis (e.g., by including perfunctory citations to one's prior work or by reciprocating citations from others) and more in line with actual (cumulative) knowledge development.

Citation classification is the basis for many scientometric analyses. The measurement of citation impact, and the mapping and visualization of scientific fields are issues of interest in scientometric research that could benefit from a qualitative distinction of citation impact. For science mapping approaches, which include co-citation and bibliographic coupling, applying our classifiers would help to reduce perfunctory influences in a more precise way compared to current practice, which is generally based on crude corrections, such as dropping certain percentages of citations before clustering. Bibliographic coupling techniques, which tend to be applied to smaller datasets compared to co-citation analyses, would provide a natural starting point to introduce ideational impact into science mapping approaches. Scientometric studies that explain the citation impact of different types of papers could use ideational impact instead of

citation counts as the dependent variable. A more nuanced analysis of this variable in terms of ideational and perfunctory impact could help increase explanatory power of scientometric models by removing biases related to perfunctory impact.

Tools for academic literature searches have been dominated by two main approaches: a keyword-based search applying specific subject terms to reduce literature to a topic and a citation-based search starting from a key document to identify related literature citing the key document. This latter way of using citations as a search tool (Hassan and Loebbecke 2017) has become increasingly popular due to its efficiency and the availability of readily accessible citation indices such as those provided by Google Scholar and Web of Science. Future research could build on our classifiers to implement new search tools that are based on ideational impact and enable scholars to find more relevant research.

Conclusion

The aim of this study was to develop an automated approach toward the classification of ideational impact of IS RAs and evaluate the model's effectiveness. With the proposed ensemble machine learning classifier, the evaluation results provide evidence for an effective and scalable classification approach that can help identifying ideational impact. We show how NLP-based features on the syntactic, semantic, and contextual level from both citing and cited papers can be used to enhance the classification performance of ideational impact. The approach presents a technical solution to the ideational impact classification problem diverging from existing approaches because it does not depend on manual qualitative analysis, and therefore allows reproducible, large-scale identification of cumulative knowledge development. Success in developing approaches able to address the task of reliable, large-scale identification of ideational impact for papers, research streams or even entire disciplines – as this paper found possible for RAs in the IS business value domain – has the potential to make to process of tracking cumulative knowledge development reproducible and transparent. Hence, this study could help achieve what Merton suggested in his early works on the sociology of science: "Having access to cumulative opportunity for scholarly work is one thing; seizing that opportunity and putting it to effective use is quite another" (Merton and Gaston 1977, p. 93).

Acknowledgements

We are grateful for the feedback provided by the track chairs, associate editor and the anonymous reviewers, which helped us to clarify the classification approach and the contributions. We would also like to thank Kai Larsen, Emrah Yasasin and Richard Schuster for their feedback on earlier versions of this paper. The research is supported by a grant of the German Science Foundation (DFG) for the research project "Epistemological Advances Through Qualitative Literature Reviews in Information Systems Research" (EPIQUALIS) (http://gepris.dfg.de/gepris/projekt/315925033?language=en).

References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker, J. F. 2010. "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly*, (34:3), pp. 435–461.

Athar, A. 2011. "Sentiment Analysis of Citations Using Sentence Structure-based Features," in *Proceedings of the 49th Meeting of the Association for Computational Linguistics*, Portland, OR.

Bartis, E., and Mitev, N. 2008. "A Multiple Narrative Approach to Information Systems Failure: A Successful System That Failed," *European Journal of Information Systems*, (17:2), pp. 112–124.

Bornmann, L., and Daniel, H.-D. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behavior," *Journal of Documentation*, (64:1), pp. 45–80.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. 1984. *Classification and Regression Trees*, New York, NY: CRC Press.

Cuellar, M., Takeda, H., and Truex, D. 2016. "Assessing the Academic Influence of a Scholarly Paper," in *Proceedings of the 6th Pre-ICIS SIGPhil Workshop*, Dublin, Ireland.

Debortoli, S., Müller, O., Junglas, I., and Brocke, J. vom. 2016. "Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems*, (39), pp. 110–135.

Dehning, B., Richardson, V. J., Urbaczewski, A., and Wells, J. D. 2004. "Reexamining the Value Relevance of E-commerce Initiatives," *Journal of Management Information Systems*, (21:1), pp. 55–82.

Dong, C., and Schäfer, U. 2011. "Ensemble-style Self-training on Citation Classification.," in *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand.

Fawcett, T. 2006. "An Introduction to ROC Analysis," *Pattern Recognition Letters*, (27:8), pp. 861–874.

Garfield, E., and Merton, R. K. 1979. *Citation Indexing, Its Theory and Application in Science, Technology, and Humanities*, New York, NY: Wiley.

Grover, V., Raman, R., and Stubblefield, A. 2013. "What Affects Citation Counts in MIS Research Articles? An Empirical Investigation," *Communications of the Association for Information Systems*, (34), pp. 1435–1456.

Guo, C., Yu, Y., Sanjari, A., and Liu, X. 2014. "Citation Role Labeling Via Local, Pairwise, and Global Features," in *Proceedings of the 77th Annual Meeting of the American Society for Information Science and Technology*, Seattle, WA.

Han, J., Pei, J., and Kamber, M. 2011. *Data Mining: Concepts and Techniques*, Waltham, MA: Elsevier.

Hansen, S., Lyytinen, K., and Markus, M. L. 2006. "The Legacy of 'Power and Politics' in Disciplinary Discourse: A Citation Analysis," in *Proceedings of the 27th International Conference on Information Systems*, Milwaukee, WI.

Hassan, N. R., and Loebbecke, C. 2017. "Engaging Scientometrics in Information Systems," *Journal of Information Technology*, (32:1), pp. 85–109.

Holsti, O. 1969. *Content Analysis for the Social Sciences and Humanities*, Reading, MA: A-W.

Jackson, D. N., and Rushton, J. 1987. *Scientific Excellence: Origins and Assessment*, New York, NY: SAGE.

Jochim, C., and Schütze, H. 2012. "Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme," in *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India.

Judge, T. A., Cable, D. M., Colbert, A. E., and Rynes, S. L. 2007. "What Causes a Management Article to Be Cited — Article, Author, or Journal?," *Academy of Management Journal*, (50:3), pp. 491–506.

Jurafsky, D., and Martin, J. H. 2009. *Speech and Language Processing*, Upper Saddle River, NJ: Pearson.

Keen, P. G. W. 1980. "MIS Research: Reference Disciplines and a Cumulative Tradition," in *Proceedings of the 1st International Conference on Information Systems*, Philadelphia, PA.

Kohli, R., and Devaraj, S. 2003. "Measuring Information Technology Payoff: A Meta-Analysis of Structural Variables in Firm-Level Empirical Research," *Information Systems Research*, (14:2), pp. 127–145.

Larsen, K., and Bong, C. H. 2016. "A Tool for Addressing Construct Identity in Literature Reviews and Metaanalyses," *MIS Quarterly*, (40:3), pp. 529–551.

Larsen, K. R., Monarchi, D. E., Hovorka, D. S., and Bailey, C. N. 2008. "Analyzing Unstructured Text Data: Using Latent Categorization to Identify Intellectual Communities in Information Systems," *Decision Support Systems*, (45:4), pp. 884–896.

Lu, C., Ding, Y., Schnaars, M., and Zhang, C. 2017. "Understanding the Impact Change of a Highly Cited Article: A Content-based Citation Analysis," *Scientometrics*, (112:2), pp. 927–945.

MacRoberts, M. H., and MacRoberts, B. R. 1989. "Problems of Citation Analysis: A Critical Review," *Journal of the American Society for Information Science*, (40:5), pp. 342–349.

Martens, D., and Provost, F. 2014. "Explaining Data-driven Document Classifications," *MIS Quarterly*, (38:1), pp. 73–99.

Merton, R. K., and Gaston, J. 1977. *The Sociology of Science in Europe*, Chicago, IL: Southern Illinois University Press.

Mingers, J., and Xu, F. 2010. "The Drivers of Citations in Management Science Journals," *European Journal of Operational Research*, (205:2), pp. 422–430.

Mithas, S., Tafti, A., Bardhan, I., and Goh, J. M. 2012. "Information Technology and Firm Profitability: Mechanisms and Empirical Evidence," *MIS Quarterly*, (36:1), pp. 205–224.

Mitkov, R. 2005. *The Oxford Handbook of Computational Linguistics*, Oxford, UK: Oxford University.

Moravcsik, M. J., and Murugesan, P. 1975. "Some Results on the Function and Quality of Citations," *Social Studies of Science*, (5:1), pp. 86–92.

Neuendorf, K. A. 2002. *The Content Analysis Guidebook*, Thousand Oaks, CA: SAGE Publications.

Nevo, S., and Wade, M. 2011. "Firm-level Benefits of IT-enabled Resources: A Conceptual Extension and an Empirical Assessment," *The Journal of Strategic Information Systems*, (20:4), pp. 403–418.

Nicolaisen, J. 2007. "Citation Analysis," *Annual Review of Information Science and Technology*, (41:1), pp. 609–641.

Pappas, N., and Popescu-Belis, A. 2016. "Human Versus Machine Attention in Document Classification: A Dataset with Crowdsourced Annotations," in *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media*, Austin, TX, pp. 94–100.

Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. 2015. "Synthesizing Information Systems Knowledge: A Typology of Literature Reviews," *Information & Management*, (52:2), pp. 183–199.

Pham, S. B., and Hoffmann, A. 2003. "A New Approach for Scientific Citation Classification Using Cue Phrases," in *Proceedings of the 16th Australasian Joint Conference on Advances in Artificial Intelligence*, Perth, Australia, pp. 759–771.

Rivard, S. 2014. "Editor's Comments: The Ions of Theory Construction," *MIS Quarterly*, (38:2), pp. iii–xiv.

Rowe, F. 2014. "What Literature Review is Not: Diversity, Boundaries and Recommendations," *European Journal of Information Systems*, (23:3), pp. 241–255.

Sabherwal, R., and Jeyaraj, A. 2015. "Information Technology Impacts on Firm Performance: An Extension of Kohli and Devaraj (2003)," *MIS Quarterly*, (39:4), pp. 809–836.

Schryen, G. 2013. "Revisiting IS Business Value Research: What We Already Know, What We Still Need to Know, and How We Can Get There," *European Journal of Information Systems*, (22:2), pp. 139–169.

Schryen, G., Wagner, G., and Benlian, A. 2015. "Theory of Knowledge for Literature Reviews: An Epistemological Model, Taxonomy and Empirical Analysis of IS Literature," in *Proceedings of the 36th International Conference on Information Systems*, Fort Worth, TX.

Serenko, A., and Dumay, J. 2015. "Citation Classics Published in Knowledge Management Journals. Part II: Studying Research Trends and Discovering the Google Scholar Effect," *Journal of Knowledge Management*, (19:6), pp. 1335–1355.

Sidorova, A., Evangelopoulos, N., Valacich, J. S., and Ramakrishnan, T. 2008. "Uncovering the Intellectual Core of the Information Systems Discipline," *MIS Quarterly*, (32:3), pp. 467–482.

Sinha, A. P., and May, J. H. 2004. "Evaluating and Tuning Predictive Data Mining Models Using Receiver Operating Characteristic Curves," *Journal of Management Information Systems*, (21:3), pp. 249–280.

Small, H. G. 1978. "Cited Documents As Concept Symbols," *Social Studies of Science*, (8:3), pp. 327–340.

Smith, L. C. 1981. "Citation Analysis," *Library Trends*, (30:1), pp. 83–106.

Stremersch, S., Camacho, N., Vanneste, S., and Verniers, I. 2015. "Unraveling Scientific Impact: Citation Types in Marketing Journals," *International Journal of Research in Marketing*, (32:1), pp. 64–77.

Tahamtan, I., Afshar, A. S., and Ahamdzadeh, K. 2016. "Factors Affecting Number of Citations: A Comprehensive Review of the Literature," *Scientometrics*, (107:3), Springer, pp. 1195–1225.

Takeda, H., Cuellar, M., Truex, D., and Vidgen, R. 2011. "Networks of Innovation in IS Research: An Exploration of the Relationship between Co-authorship Networks and H-family Indices," in *Proceedings of the 19th European Conference on Information Systems*, Helsinki, Finland.

Teufel, S., Siddharthan, A., and Tidhar, D. 2009. "An Annotation Scheme for Citation Function," in *Proceedings of the 7th SIGDIAL Workshop on Discourse and Dialogue*, pp. 80–87.

Truex, D., Cuellar, M., and Takeda, H. 2009. "Assessing Scholarly Influence: Using the Hirsch Indices to Reframe the Discourse," *Journal of the Association for Information Systems*, (10:7), pp. 560–594.

Valenzuela, M., Ha, V., and Etzioni, O. 2015. "Identifying Meaningful Citations," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX.

Wagner, G., Prester, J., Roche, M., Benlian, A., and Schryen, G. 2016. "Factors Affecting the Scientific Impact of Literature Reviews: A Scientometric Study," in *Proceedings of the 37th International Conference on Information Systems*, Dublin, Ireland.

Wan, X., and Liu, F. 2014. "Are All Literature Citations Equally Important? Automatic Citation Strength Estimation and Its Applications," *Journal of the Association for Information Science and Technology*, (65:9), pp. 1929–1938.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly*, (26:2), pp. xiii–xxiii.

Wuchty, S., Jones, B. F., and Uzzi, B. 2007. "The Increasing Dominance of Teams in Production of Knowledge," *Science*, (316:5827), pp. 1036–1039.

Xue, Y., Liang, H., and Wu, L. 2011. "Punishment, Justice, and Compliance in Mandatory IT Settings," *Information Systems Research*, (22:2), pp. 400–414.