Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6777–6785
Marseille, 11–16 May 2020

Beyond Citations: Corpus-based Methods for Detecting the Impact of Research Outcomes on Society

Rezvaneh Rezapour[§], Jutta Bopp[†], Norman Fiedler[†], Diana Steffen[†], Andreas Witt[†], Jana Diesner[§]
§University of Illinois at Urbana–Champaign, USA
{rezapou2,jdiesner}@illinois.edu
†Leibniz-Institut für Deutsche Sprache, Germany
{bopp,fiedler,steffen,witt}@ids-mannheim.de

Abstract

This paper proposes, implements and evaluates a novel, corpus-based approach for identifying categories indicative of the impact of research via a deductive (top-down, from theory to data) and an inductive (bottom-up, from data to theory) approach. The resulting categorization schemes differ in substance. Research outcomes are typically assessed by using bibliometric methods, such as citation counts and patterns, or alternative metrics, such as references to research in the media. Shortcomings with these methods are their inability to identify impact of research beyond academia (bibliometrics) and considering text-based impact indicators beyond those that capture attention (altmetrics). We address these limitations by leveraging a mixed-methods approach for eliciting impact categories from experts, project personnel (deductive) and texts (inductive). Using these categories, we label a corpus of project reports per category schema, and apply supervised machine learning to infer these categories from project reports. The classification results show that we can predict deductively and inductively derived impact categories with 76.39% and 78.81% accuracy (F1-score), respectively. Our approach can complement solutions from bibliometrics and scientometrics for assessing the impact of research and studying the scope and types of advancements transferred from academia to society.

Keywords: impact assessment, natural language processing, machine learning, corpus analysis, deductive and inductive category detection

1. Introduction

National and independent organizations and foundations have been providing researchers across domains with opportunities in the form of funding and infrastructure to advance knowledge and discovery. The premise with research funding is that the return of investment, i.e., research outcomes, will benefit people, the economy, or the environment, among other beneficiaries (Bornmann and Daniel, 2005). How do we know if this goal has been achieved? Traditionally, the performance of researchers and the impact of research outputs have been assessed by capturing and analyzing citations of publications. The field of informetrics has been developing various such metrics, which can be factors considered when allocating research funding or evaluating research (Aksnes et al., 2019). In addition to that, funders, scholars and society have become interested in measuring traceable effects of research outcomes beyond their reach within academia[1] (Parker and Van Teijlingen, 2012).

In response to these needs and given the growth of research being mentioned on online platforms and social media, additional metrics that capture attention to scholarly work have been developed[2] (Priem et al., 2012). These metrics are being used to supplement citation counts and expanding the types of impact of research that are considered for assessments (Piwowar, 2013; Bornmann, 2015; Pulido et al., 2018; Subramanyam, 1983; Swanson et al., 2006; Smalheiser and Torvik, 2008). Also, the field of science of science has developed and studied additional indicators of the impact of scientists and their work, such as awards and promotions (Fortunato et al., 2018). The mentioned metrics provide a scalable and quantifiable approach for measuring different types of impact of scholarly work. In addition to that, there is an increasing interest in adequately representing the broader influence of research on society, and how new knowledge and ideas get transferred from academia to the public.

In this context, the lack of transparency and interpretability of the impact of scholarly work are key limiting factors, especially for assessing publicly funded research. The general public increasingly asks for cross-references to the allocation of taxpayers' money. However, the intellectual, factual and material access of the public to research can be a challenging task due to a lack of language consistency across academic domains, domain-specific terminology, limited open access resources, and publishers' paywalls. Even though the public has a right to benefit from science (Wyndham and Vitullo, 2018; American Association for Advancement of Science, nd), and many universities aim to benefit their communities and the public (Tsey, 2019), current system of disseminating research outcomes and evaluating their impact are not necessarily accessible and understandable to non-academics (Bornmann, 2012), which further limits scholars in producing science for public good (Berendt, 2019).

To meet the need for advancing the measurement of societal and economic influences of science (Bornmann, 2013; Bornmann, 2012), we present a solution for bridging the gap between the scholarly domain and society by (i) introducing, implementing and evaluating text-based indicators of scholarly impact on the real world, and (ii) developing and evaluating computational methods and resulting models for extracting indicators of social and economic impacts of (publicly funded) research from project reports. By doing so, this work contributes to (a) enabling researchers

[1] https://www.ref.ac.uk/2014/pubs/2011-02/
[2] https://www.altmetric.com/

to assess the impact of their work and providing meaningful analyses of their contributions to society, (b) improving transparency over the return of public investments in research to society, and (c) curtailing opportunities for adversarial attacks that are taking advantage of the citation-count based system via gaming bibliometric scores to boost papers and authors (Aksnes et al., 2019).

Researchers and funding agencies have been trying to address these issues by developing taxonomies and frameworks that aim at better understanding the social and economic impact of projects in fields such as healthcare, agriculture, and environmental studies (Bornmann, 2013; Bornmann, 2017; Tsey, 2019; Tsey et al., 2019; Wolf et al., 2013; Heyeres et al., 2019; Greenhalgh et al., 2016; Vanclay, 2003). What is still missing is a generalizable impact framework that is applicable across domains. Moreover, due to a lack of standardized structure and language used to write up research results and reports, studying the impact of research outcomes requires human expertise as well as advanced technical solutions to go through large sets of texts to extract the relevant information. Manual evaluation is limited by the large and growing number of research papers, and suitable automated solutions are yet to be developed.

To address these shortcomings, we present a novel framework that considers two computational, human-in-the-loop approaches: a deductive (top-down, from theory to data) one, and an inductive (bottom-up, from data to theory) one, and apply them to data to develop two impact category schemes. To implement and test both approaches, we use a mixed-methods strategy, and contrast survey-based methods with text-mining techniques for impact assessment. For the deductive approach, impact categories were derived from prior research and expert knowledge on academic impact assessment. We then interviewed researchers and principal investigators of grant-funded work to assign the identified categories to the projects considered herein. For the inductive approach, we postulate that project reports may explicitly or implicitly express actual or potential implications of the presented research. To test this assumption, we used close reading as a technique to extract various types of impact, such as influence on people's well-being and impact on societal awareness, from project reports (Table 1). We refer to these empirically grounded categories as "anticipated" impact since these indicator phrases were stated in reports, which may precede the transfer of science to society. Using the deductive and the inductive approach, we labeled a set of project reports, which resulted in two different annotations of the same dataset. We then leveraged methods from natural language processing (NLP) and machine learning to build a model per annotation or approach for predicting each set of impact categories, tested the performance, and compared the outcomes.

Analyzing the results from the deductive approach shows that, overall, projects address multiple types of impact, and that the majority of funded projects aim for technical and economic impact. This may be due to the considered domain, which we elaborate on in the data section. In addition, the results of the inductive approach show that the considered science projects aim for improving knowledge and having an impact on society and the public. Researchers indicate societal impact, and discuss potential benefits of their work for educational purposes and raising awareness in society about the outcomes of their work. The combination of the deductive and the inductive approach shows that funded projects often focus on making an impact on scientific domains, and discuss the outcomes of their work in the form of products, publications, or guidelines. The results from testing our classifiers show that one can automatically distinguish the impact categories developed in the deductive and inductive approaches with 76.39% and 78.81% accuracy (F1 score), respectively. We aim to use these models in future work to build a computational impact system that helps to translate different types of research impact into information that is accessible and understandable for the general public.

The approaches and insights discussed in this paper contribute to research in the areas of impact assessment and science of science. Our mixed-methods approach can be used as a complementary solution to bibliometric methods. In addition, we will release our annotated datasets to provide researchers interested in impact assessment with opportunities for testing prediction solutions and conceptual ideas.

2. Literature Review

The focus of Social Impact Assessment (SIA) is to identify, understand and estimate the influence or consequences of actions and objects on individuals, groups and communities, or society (Latané, 1981). Analyzing impact can facilitate decision-making processes and minimize risks of investments.

Impact assessment has been studied and practiced in domains such as environmental studies (Becker, 2001; Becker et al., 2003; Vanclay, 2006), economics (Shmueli and others, 2010), psychology (Latané, 1981), and political studies (Grimmer and Stewart, 2013), to name a few. In environmental studies, for example, SIA is focused on studying the consequences of planned or unplanned events, and monitoring and managing these consequences (Vanclay, 2006). Moreover, after the anticipated consequences of a project have been identified, researchers share their plans with the public and can change a project according to suggestions and feedback they receive. In political science or economics, researchers mainly focus on causal explanations through regression analysis and statistical inferences to detect relations between stimuli (sources) and social impact to estimate the magnitude of effects. It is worth noting that often the process of causal detection and explanation is of higher interest than measuring the impact itself (Grimmer and Stewart, 2013; Shmueli and others, 2010).

In recent years, (philanthropic) foundations had begun to request traceable evidence and reports from their grantees to analyze the social return on investment (Chattoo and Das, 2014; Barrett and Leddy, 2008; Clark and Abrash, 2011; Diesner et al., 2016). To develop reliable, feasible, cost-efficient and acceptable solutions to impact assessment, some foundations have been collaborating with academia and scholars, which has resulted in guidelines, frameworks and methods for assessing the impact of funded

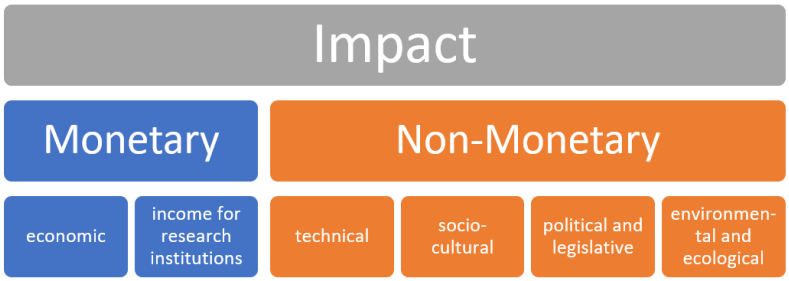

Figure 1: Classification schema for deductively derived model using external impact categories

projects (Rezapour and Diesner, 2017; Blakley et al., 2016; Napoli, 2014; Diesner et al., 2014; Diesner and Rezapour, 2015; Witt et al., 2018; Aufderheide, 2015). In academia, the impact and quality of research have been measured via bibliometric techniques, such as citation counts and related metrics such as the h-index, and peer review assessment, respectively (Bornmann and Daniel, 2005; Hirsch, 2005; Van Raan, 2004; Van Raan, 1996). While these methods can provide a reliable understanding of the intellectual influence of scientific output, they fall short of highlighting the social and economic impact of funded research, and how much of the produced knowledge is transferred to the public. This has been remedied by more recent efforts, such as the altmetrics movement, that consider the impact of research beyond academia, for example, by analyzing mentions of research in traditional and social media, or tracking the sharing and reuse of resources and data (Piwowar, 2013; Taylor, 2013). Furthermore, several countries and organizations have started to develop frameworks and guidelines for assessing the social and economic impact of funded research (Bornmann, 2013; Bornmann, 2017; Tsey, 2019; Tsey et al., 2019; Wolf et al., 2013; Heyeres et al., 2019; Greenhalgh et al., 2016). For example, the 'Payback Framework" is used in health-related studies to assess the impact of research on knowledge, future work, policy, health-related applications, and economic benefits (Greenhalgh et al., 2016; Heyeres et al., 2019; Gomes and Stavropoulou, 2019). The CAHS framework (Canadian Academy of Health Services) aims to measure the advancement of knowledge, capacity-building, and economic and social benefits such as commercialization, cultural outcomes, socioeconomic implications, and the public understanding of science. These frameworks leverage individual and focus group interviews, bibliometrics, case studies, and archival data for assessing impact (Frank et al., 2009).

While prior work on impact assessment provides substantial frameworks, taxonomies and insights for capturing the influence of scientific work on the research community, they lack in scalability since executing them for practical assessment studies is costly, time consuming, and in some cases biased (Greenhalgh et al., 2016; Tsey, 2019; Tsey et al., 2019; Wolf et al., 2013). This paper presents a comparatively comprehensive approach to assessing the impact of research beyond academia by analyzing project reports with text-based analysis and conducting interviews and surveys with project personnel in order to train prediction models of impact. Our methods and findings can help in better understanding the direct and indirect impact of research outcomes on society.

3. Data

We work with a corpus of final reports of publicly funded research projects provided by the Leibniz Information Centre for Science and Technology (TIB[3]). The TIB serves as the German National Library of Science and Technology. Given the large number of digital reports (around 75k) available from to the TIB, we restricted our corpus to one scientific domain, i.e., *mobility*. This is a multidisciplinary field that brings together different disciplines, including engineering, urban studies and social science, which we consider beneficial to this project as it may increase the generalizability of our work. To extract the reports to be considered in our corpus, we used the following criteria:

- Report(s) that are digitally available in the TIB library (PDFs and metadata)
- Project type (based on project meta-data): technology and promotion of innovation
- Project completion: between 2005 and 2015
- Publicly funded, collaborative projects with two to ten partners
- At least one academic project partner

Given these search parameters, our corpus consists of 91 projects with a total of 391 individual reports (each project can consist of more than one report, for example, when there is a final joint report and individual reports from different project partners). We converted the reports to plain text and removed all non-textual data such as pictures, complex mathematical typesetting, and tables. The reports are all in German.

4. Defining Impact and Data Annotation

Information products can affect people or society in various ways. While some may impact people directly, others may take years to show their influence or impact people in an indirect form (Rezapour and Diesner, 2017). To analyze the impact of funded research, we used two approaches: a deductive (top-down, from theory to data) and an inductive (bottom-up, from data to theory) one. In the following sections, we explain the process for developing impact categories and annotating our corpus.

[3]https://www.tib.eu/en/

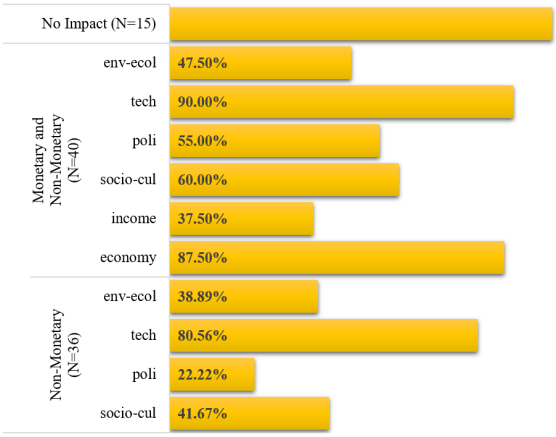

Figure 2: Distribution of deductively derived impact categories and sub-categories

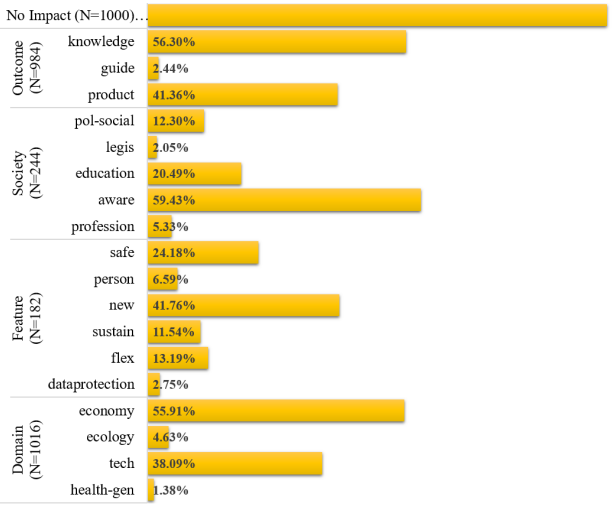

Figure 3: Distribution of inductively derived impact categories and sub-categories

4.1. Deductive: Top-down, from Theory to Data

To define data-driven and meaningful impact categories for our corpus, we studied prior frameworks, and solicited input from domain experts in the area of management with a focus on innovation and transfer management (Witt et al., 2018). Based on this input and our discussions, we defined six impact categories of research projects:

- Economic impact (**economy**): refers to the use of research results in the private sector, e.g., the development of a business model.
- Income impact (**income**): refers to additional income for research institutions, e.g., selling licenses or research contracts.
- Technical impact (**tech**): refers to technologies that are used outside of the original project, e.g., prototype development or process development.
- Socio-cultural impact (**socio-cul**): occurs when a project influences societal groups or institutions like schools, local authorities, foundations, or clubs, also includes activities such as starting a grass-root initiative.
- Political impact (**poli**): refers to using the project results in political or jurisdictional contexts, e.g., contributions to a new law or informing political advice.
- Environmental and ecological impact (**env-ecol**): refers to changes of ecological or environmental aspects, e.g., environmental reports or weather data collection.

We then associated these six categories with broader ones, i.e., “Monetary Impact” (economic or income impact), “Non-monetary Impact” (technical, socio-cultural, political and legislative, and environmental and ecological impact), “Monetary and Non-monetary Impact”, and ”No Impact” (with respect to this project) (Figure 1). We categorize a project as having no impact if (1) it does not show any relevant contribution other than purely scientific impact (which is still significant impact, but not considered for the scope of this project), or (2) the interviewed project member had no sufficient knowledge about the project outcomes after the project had officially ended.

4.1.1. Top-down Data Annotation

In order to label the projects in our corpus, we first identified the principal investigator or a relevant project member of each project. We explained the purpose of our study to them and asked for their permission to conduct an interview regarding the project. The project members were informed about our compliance with data protection regulations and the anonymity of the knowledge gained from the data collection.

Upon their agreement, the project members completed a questionnaire regarding various types and aspects of the impact of their project. Moreover, we asked them if their project had any influence on (different areas of) society, about the form of achieved impact, who was involved in that, which of the six categories described above best represents their achieved impact, and if the project generated any income for their research institution. We then mapped each project to one impact category (“Monetary Impact”, “Non-monetary Impact”, “Monetary and Non-monetary Impact”, and ”No Impact”). The annotations were on the project level, and all reports associated with one project were given the same category. Figure 2 shows the distribution of the categories and sub-categories in our corpus.

4.2. Inductive: Bottom-up, from Data to Theory

This process relies only on text data, i.e., the reports in our corpus, and disregards the labels assigned as described above. In a report, impact may be represented by describing methods and routines implemented for a project, or by the impact that authors anticipated when writing their final reports. To identify text-based evidence of impact, we trained and asked three human annotators to closely read a set of reports that was randomly selected from our corpus, mark up sentences or sections that indicate any type of impact (no pre-defined categories given (free recall), and the annotators did not know the categories from the top-down approach), and provide a label for the impact types related to the extracted text). Since the reports are in German, we translated the selected sample into English, and asked three different human annotators (English speakers) to perform

Impact		Impact Information
Impact on Domain, Area or Field	Definition	Impact on domain (i.e., application)
	Sub-categories	1) Impact on economy **(economy)**: both macroeconomic aspects (economic developments) and microeconomic aspects (marketing, business model, costs, sales, strategies) 2) Impact on ecology **(ecology)**: energy turnaround, environmental protection, sustainability, climate protection 3) Impact on (general) health **(health-gen)**: more road safety, fewer accidents, more mobility for old people, fewer depressions, fewer pollutants, fewer respiratory diseases 4) Impact on technology **(tech)** (e.g., electro-mobility, autonomous driving, high performance computing/computer technology)
Impact on Society and Public Sphere	Definition	Impact on social or public circumstances, processes and institutions; changes in social principles (e.g., common language, standards, regulations for deviant behavior, and understanding)
	Sub-categories	1) Legal and legislative impact **(legis)** (e.g., benefit of the new product on legislations and rules, changes in laws and legislations, etc.) 2) Impact on public health services **(health service)** (in contrast to general health in the 1st category) (e.g., Vaccination campaign in emergency situations etc.) 3) Impact on public education/general education **(education)** (e.g., new master program; inclusion etc.) 4) Impact on professional world **(profession)** (e.g., employment, unemployment, new professions emerge, old professions disappear, job profiles change etc.) 5) Impact on political/social issues **(pol-social)** (e.g., climate change, refugees/migration, religious persecution, etc.) 6) Impact on awareness/perception **(aware)** (e.g., events (open days, exhibitions, press conferences), newspaper articles, radio/television contributions, social media, campaigns etc.)
Impact on Outcomes or Products	Definition	Impact outcome represents the final result of a project (indicated) in reports
	Sub-categories	1) Impact in the form of real products or prototypes **(product)** (physical and non-physical): e.g., iphone, autonomous car; apps, online platforms, eCourses, eBooks; specific data, e.g., lists of email addresses sold as products by brokers services 2) Knowledge-based impact **(know)**: e.g., methods: research methods, learning and teaching methods, algorithms concepts, models, data (if mentioned in the report but not further specified) innovative, faster and more efficient technical procedures (mostly result in patents) 3) Impact in form of guidelines **(guide)**: e.g., uniform standard for mobile phone stickers 4) Other impact **(other)**: Possible new relevant sub-categories that have not yet been considered
Impactful Features or Characteristics of Products, Services, and Public Goods	Definition	Various characteristics of outcomes and impacts
	Sub-categories	1) Novelty **(new)**: a truly new and innovative result. We don't consider a feature novel if it is discussing optimization or improvement of existing platforms or methods 2) Safety **(safe)**: the outcome offers/supports (more) safety, e.g., Road safety, users' safety, general safety 3) (Data) protection **(dataprotection)**: the outcome ensures (more) (data) protection and privacy 4) Sustainability **(sustain)**: the outcome is sustainable 5) Flexibility **(flex)**: the result allows more flexibility 6) Personalization **(person)**: the result can be personalized 7) Other impact **(other)**: possible new relevant subcategories that have not yet been considered

Table 1: Inductively derived impact categories

the same procedure. None of the annotators had any prior knowledge about the projects or their impact before the annotation task. Once the annotators had completed their task, we synthesized their inputs. Through consultation among our team members, we consolidated or normalized the annotators' free-recall categories, which resulted in an alternative category schema that was built bottom-up. We iteratively tested and refined this schema until we believed it to be comprehensive and meaningful for our task.

The resulting schema is shown in Table 1, and contains four main impact types: (1) "Impact on Domain, Area or Field", (2) Impact on Society and the Public Sphere", (3) "Impact on Outcomes or Products", and (4) "Impactful Features or Characteristics of Products, Services, and Public Goods". We further defined multiple sub-categories for our codebook (Table 1) to highlight the depth and magnitude of each main type, and to help non-experts in understanding and distinguishing the types we defined.

4.2.1. Bottom-up Data Annotation

To make the application of the codebook to label each project more efficient, we (1) identified common sections of the reports that address achieved or the potential impact, and marked these sections for the human coders, and, (2) for projects with multiple reports, we selected one report per each project; preferably the one with the overall results of the projects. A total of four annotators (six pairs) annotated the relevant sections of selected documents. The annotators were allowed but not encouraged to choose more than one category per section. Overall, the annotators assigned the same label to 60% of the sentences, and provide no or different labels to 40% of the sentences. The average kappa value (of all six pairs) was around 48%.

Furthermore, sentences with disagreement were adjudicated by two researchers who were not involved in the original annotation. Figure 3 provides information on the final set of labeled sentences. The resulting annotated corpus is labeled with impact categories and sub-categories. We will

publicly share this resources upon finalizing its preparation for release.

5. Feature Selection and Classification

After labeling the input corpus twice (via the deductive and inductive approach), we used the labeled texts as input to train prediction models. Since both datasets are small, we decided to use classic, feature-based machine learning algorithms (Support Vector Machines (SVM), Gaussian Naive Bayes, and Random Forest). In this paper, we only report the result of the SVM model since it achieved the highest accuracy.

To extract features, we first preprocessed the data by removing numbers, symbols, e.g., umlauts and stop words, and words that appeared in less than 5% and more than 95% of the data.

To build classifiers for both version of the labeled data, we used three sets of features: (1) lexical features (tf-idf, for which we used the vectorizer in Python's SKLearn library to extract unigram, bigrams and trigram), (2) syntactic features (Parts of Speech (POS)), and (3) domain-specific features (impact sub-categories).

We used the lexical unigram features as a baseline for both approaches, and added the rest of the features on top of that to analyze their contribution to prediction accuracy.

To extract syntactic features, we used TextBlob, German package[4] (Loria et al., 2018) to tag each word with its POS (this was done prior to data cleaning). We then counted the number of nouns, verbs, adjectives, adverbs, etc. in each entry, and added them as additional features on top of the lexical features.

For the domain-specific features, we used the impact sub-categories. For the deductive approach, these are economic, income, technical, socio-cultural, political, or environmental and ecological impact (last row in Figure 1). Each project can be associated wit one or more sub-categories. For the inductive approach, we used the categories shown in **bold** in the third columns of Table 1). We added these additional features on the top of lexical and syntactic features.

To address the skewedness of instances per category in our data, we used Synthetic Minority Over–sampling TEchnique (SMOTE) to increase the number of instances in smaller categories and balance the input data for training classifiers (Chawla et al., 2002). With the deductive model, as shown in Figure 2, the "Monetary and Non-Monetary Impact" class has the most instances (N=40). We oversampled instances in the smaller class, namely "No Impact" to balance the input data. With the inductive approach (Figure 3), we under-sampled the largest class "No impact" by randomly selecting 1000 sentences, and then used SMOTE to synthetically over-sample the small classes, namely "Feature" and "Society".

Finally, to increase the performance of classifiers and reduce the redundancy of features, we leveraged χ^2 (Chi-Square) algorithm (equation (1)) to select the top k ($300 < k < 600$) attributes for both approaches. More specifically, χ^2 is used to test whether the occurrence of a specific term and a specific class are independent of each other. Given a document D , we estimate the following quantity for each term, and rank the terms by their score:

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \quad (1)$$

In equation (1), N shows the observed frequency, and E represents the expected frequency. If a document contains term t, e_t takes the value of 1, and 0 otherwise. If the document is in class c, e_c takes the value 1. The values for both e_t and e_c will be 0 if the rule is not satisfied.

We used Python Sklearn package (Pedregosa et al., 2011) to implement the algorithms and classifiers, 5-fold cross validation to train the models, and standard metrics (precision, recall, F1, Area Under the Receiver Operating Characteristic Curve (ROC AUC)) to assess prediction accuracy.

6. Results

6.1. Data Analysis

None of the projects labeled with the deductive approach had "Monetary Impact" (Figure 2), and none was solely focused on economic or income impact. The majority of projects (43.95%) were reported to have had "Monetary and Non-Monetary Impact"; 39.56% had "Non-Monetary Impact"; and 16.48% had no monetary or non-monetary impact. Analyzing the sub-categories from the deductive approach shows that the majority of funded projects (71.4%) focused on technical impact (Figure 2). We also found that 87.5% of the projects labeled as "Monetary and Non-Monetary Impact" represent some sort of economic impact. Only 16.48% of all projects focused on increasing income in institutions. Regarding the socio-technical impact of projects, we found that 42.85% of the projects are associated with affecting societal groups or institutions. In the inductively labeled dataset, 60.23% of the annotated sentences do not carry any information related to impact. This finding is not surprising since many sentences even in impact-relevant sections of reports provide other types of information. Moreover, we find that 16.57% of the sentences refer to "Impact on Domain", 16.17% to "Impact on Outcome", 4% to "Impact on Society and Public Sphere", and around 3% discuss "Impactful Features of Products, Services or Public Goods". Analyzing the sub-categories shows that (a) 55.91% of sentences labeled as "Impact on Domain" discuss economic impact, (b) 56.30% of sentences labeled as "Impact on Outcome" focus on improving knowledge, (c) 59.43% of sentences tagged as "Impact on Society" indicate impact on awareness/perception, and (d) 41.76% of sentences discuss novel or innovative features as outcomes of their projects. Figure 3 visualizes the number of instances per sub-category in the labeled dataset. Overall, our findings show that the majority of funded projects not only aim to advance science within the realms of academia, but also aim at advancing technologies and services for society, and providing public goods and innovative products.

We next combined both label types to further analyze the relationship between inductively and deductively derived categories. As shown in Figure 4, the majority of projects

[4]https://pypi.org/project/textblob-de/

Model	Deductively derived Model				Inductively derived Model			
	P	R	F1	ROC	P	R	F1	ROC
Unigram (Baseline)	72.37	65.81	66.45	73.38	55.62	52.06	52.95	68.91
Ngram (unigram+bigram+trigram)	77.83	75.69	75.32	80.01	56.37	52.77	53.83	69.44
Ngram + POS	77.83	75.69	75.32	80.01	56.2	52.59	53.66	69.31
Ngram + POS +Sub-categories	**80.04**	**76.87**	**76.39**	**80.82**	**79.8**	**78.29**	**78.81**	**85.92**

Table 2: Result of SVM classifier for the deductively and inductively derived model, Precision (P), Recall (R), F1 Score, Area Under the Receiver Operating Characteristic Curve (ROC AUC) (values in percent)

with "Monetary and Non-Monetary Impact" (deductive category) features "Impact on Domain" (inductive category). The majority of sentences labeled as "Non-Monetary Impact" (deductive category) discusses final research outcome such as products, prototypes, methods, and guidelines (inductive category). "Impact on Society and Public Sphere" (inductive category) is primarily discussed in projects with "Monetary and Non-Monetary Impact". Interestingly, projects with no impact (no monetary or non-monetary (deductive category) also discuss inductively derived categories (impact on domain, society, outcome, or feature) least often. We also analyzed the correlation between deductively and inductively identified categories using Pearson's correlation coeffcient. We found that "Society" is most strongly y correlated with "No Impact", while "Domain" is most strongly correlated with "Monetary and Non-Monetary Impact" (p-value <0.05).

6.2. Classification

As shown in Table 2, we first created the baseline using unigrams. For the deductively labeled data, after balancing the dataset and choosing the most informative features, the baseline model achieved an F1 score of 66.45%. For the inductively labeled data, the baseline model achieved an F1 score o 52.95%. Adding the bigrams and trigrams to the baseline increased the performance by around 10% for the deductive model, and by 1% for the inductive model. Moreover, as shown in Table 2, while precision did not change with the ngram model, recall increased by a large margin. This indicates that adding words to the feature sets helped to predict or capture true positives and increasing the classifiers' sensitivity. Adding syntactic feature on top of the lexical features did not change the performance of the classifier for the deductive model, and for the inductive model,

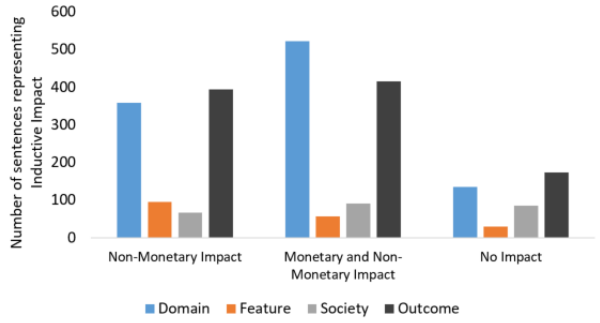

Figure 4: Distribution of deductively and inductively derived categories across projects

the decreased by 0.2%. Finally, combining the lexical, syntactic and domain-specific features increased classifier performances for both models. As shown in Table 2, the inductive model benefited the most from combining all features; achieved around 78.81% of an F1 score. The highest ROC was obtained by combining all feature sets (80.82% for deductive, 85.92% for inductive model).

7. Discussion and Conclusion

In this paper, we proposed, implemented and evaluated a novel framework and methodology for assessing the impact of funded research beyond academia. The main objectives of this work are to (1) introduce a new, corpus-based framework to supplement prior impact assessment frameworks, and (2) develop and contrast new computational methods for assessing the impact of funded research. Our novel framework consists of two approaches: a deductive (top-down, from theory to data) and an inductive (bottom-up, from data to theory) way of identifying types of real-world impact that research projects may have.

Using the deductive approach, we derived impact categories from prior work and input from experts on common and verifiable indicators of impact (Figure 1), and annotated the projects by interviewing the project members. Using the inductive approach, we extracted impact categories from final project reports by identifying text-based indicators of impact through close reading. This also resulted in a novel yet different impact scheme (Table 1), which we implemented in a codebook and used that to hand-label the texts. Both annotated corpora were then used for supervised, feature-based learning.

Overall, our results from the deductive approach show that reports of funded projects address (potential for) both societal and economic impact. The results of the interviews revealed that the majority of funded projects from the domain of mobility aim at technical and economic impact (Figure 2).

The results of our bottom-up approach supplement these findings, and show that researchers discuss anticipated or implemented impact of their projects on the economy and technology (Figure 3). In addition, funded projects mention societal impact, including benefits to education, improving or modifying legislation, and raising awareness in society (Figure 3). Combining the labels from the deductive and inductive approach (Figure 4) reveals that impact-relevant statements (mostly) refer to impact on domains and fields.

To address the second objective of this work, i.e., using the labeled data for prediction, we trained and built classification models. For the deductive approach, we classify projects, and for the inductive approach, we classify

sentences. Using the labeled data (one per approach), we extracted three sets of features (lexical, syntactic, and domain-specific), and trained three classifiers (with SVMs, Gaussian Naive Bayes, and Random Forest) for predicting impact categories. Our results 2 show that a combination of all three feature sets benefited prediction accuracy. With an F1 score of 76.39% and 78.81%, we were able to distinguish the deductively and inductively derived impact categories, respectively.

In summary, assessing the impact of research beyond bibliometrics and altmetrics is a challenging task. The scarceness of explicit indicators of impact of research on the real world as well as limited amounts of accessible data for evaluating research beyond scientific impact impose critical obstacles to studying the transfer of science to society. A large amount of prior publications and frameworks in this area theorize about the kind of societal impact that research can or should achieve. To provide an empirical approach, we apply and compare two methods for eliciting different types of impact of research projects from project personnel (deductive) and text data (inductive), respectively. We hope that these categories can help scholars and others to navigate the impact of science more comprehensively. Moreover, in the era of big data, where we experience an increase in scientific productivity and outputs, there is a need for scalable impact assessment solutions. Prior frameworks and methods are labor-intensive and time consuming. To the best of our knowledge, this work offers one of the first computational models for assessing the impact of research. This area is still in its early stages. We believe that with additional labeled data from other domains and a combination of frameworks and models, we can provide a valuable resource to the impact assessment community.

Our work is limited in several ways. First, our corpus is small, and we are only focusing on one domain (mobility). To address this shortcoming, we aim to increase the number of projects and domains. In addition, we did not verify the realization of intended impact, and treated stated intent, actual outcomes, and self-reported data as the same type of evidence. We hope to gain access to more information related to these projects to differentiate impact types for further assessment. In our future work, we aim to expand our assessment work by comparing our findings to those obtained by using commonly used bibliometric and alternative metrics. For now, we present a solution that is capable of assessing the impact of funded research projects, and provide the first study of this kind that uses German text data. We believe that consolidating these approaches will enable researchers and others to effectively assess the social impact of funded research.

8. Acknowledgements

This work is part of the project "TextTransfer - Corpus-based detection of secondary usage of scientific publications", which is funded by by the German Federal Ministry of Education and Research (BMBF) under the funding code 01IO1634. The sole responsibility for the content of this publication lies with the authors. We express our gratitude to the annotators for their tremendous help and input throughout this project.

9. Bibliographical References

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1):1–17.

American Association for Advancement of Science. (nd). Right to science.

Aufderheide, P. (2015). Conversations about impact in documentary: Beyond fear and loathing. *CineAction*, 97:2016.

Barrett, D. and Leddy, S. (2008). Assessing creative media's social impact. *The Fledgling Fund*.

Becker, D. R., Harris, C. C., McLaughlin, W. J., and Nielsen, E. A. (2003). A participatory approach to social impact assessment: the interactive community forum. *Environmental Impact Assessment Review*, 23(3):367–382.

Becker, H. A. (2001). Social impact assessment. *European Journal of Operational Research*, 128(2):311–321.

Berendt, B. (2019). AI for the common good?! pitfalls, challenges, and ethics pen-testing. *PALADYN, Journal of Behavioral Robotics*, 10(1):44–65.

Blakley, J., Huang, G., Nahm, S., and Shin, H. (2016). Changing appetites & changing minds: Measuring the impact of" food, inc.". *The USC Annenberg Norman Lear Center*.

Bornmann, L. and Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392.

Bornmann, L. (2012). Measuring the societal impact of research: research is less and less assessed on scientific impact aloneâwe should aim to quantify the increasingly important contributions of science to society. *EMBO reports*, 13(8):673–676.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? a literature survey. *Journal of the American Society for Information Science and Technology*, 64(2):217–233.

Bornmann, L. (2015). Usefulness of altmetrics for measuring the broader impact of research: A case study using data from plos and f1000prime. *Aslib Journal of Information Management*, 67(3):305–319.

Bornmann, L. (2017). Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, 73(5):775–787.

Chattoo, C. B. and Das, A. (2014). Assessing the social impact of issues-focused documentaries: Research methods & future considerations. *Center for Media and Social Impact, School of Communication at American University*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Clark, J. and Abrash, B. (2011). Social justice documentary: Designing for impact. *Center for Social Media*.

Diesner, J. and Rezapour, R. (2015). Social computing for impact assessment of social change projects.

In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 34–43. Springer.

Diesner, J., Kim, J., and Pak, S. (2014). Computational impact assessment of social justice documentaries. *Journal of Electronic Publishing*, 17(3).

Diesner, J., Rezapour, R., and Jiang, M. (2016). Assessing public awareness of social justice documentary films based on news coverage versus social media. *IConference Proceedings*.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379):eaao0185.

Frank, C., Battista, R., Butler, L., et al. (2009). Making an impact: a preferred framework and indicators to measure returns on investment in health research. *Ottawa, ON: Canadian Academy of Health*.

Gomes, D. and Stavropoulou, C. (2019). The impact generated by publicly and charity-funded research in the united kingdom: a systematic literature review. *Health research policy and systems*, 17(1):22.

Greenhalgh, T., Raftery, J., Hanney, S., and Glover, M. (2016). Research impact: a narrative review. *BMC medicine*, 14(78):1–16.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Heyeres, M., Tsey, K., Yang, Y., Yan, L., and Jiang, H. (2019). The characteristics and reporting quality of research impact case studies: A systematic review. *Evaluation and program planning*, 73:10–23.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Latané, B. (1981). The psychology of social impact. *American psychologist*, 36(4):343.

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., et al. (2018). Textblob: simplified text processing.

Napoli, P. M. (2014). *Measuring media impact: An overview of the field*. Norman Lear Center, USC Annenberg School for Communication & Journalism.

Parker, J. and Van Teijlingen, E. (2012). The research excellence framework (REF): Assessing the impact of social work research on society. *Practice*, 24(1):41–52.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830.

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, 493(7431):159.

Priem, J., Groth, P., and Taraborelli, D. (2012). The altmetrics collection. *PloS one*, 7(11).

Pulido, C. M., Redondo-Sama, G., Sordé-Martí, T., and Flecha, R. (2018). Social impact in social media: A new method to evaluate the social impact of research. *PloS one*, 13(8):e0203117.

Rezapour, R. and Diesner, J. (2017). Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1419–1431. ACM.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.

Smalheiser, N. R. and Torvik, V. I. (2008). The place of literature-based discovery in contemporary scientific practice. In *Literature-based discovery*, pages 13–22. Springer.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science*, 6(1):33–38.

Swanson, D. R., Smalheiser, N. R., and Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American Society for Information Science and Technology*, 57(11):1427–1439.

Taylor, M. (2013). Exploring the boundaries: How altmetrics can expand our vision of scholarly communication and social impact. *Information Standards Quarterly*, 25(2):27–32.

Tsey, K., Onnis, L.-a., Whiteside, M., McCalman, J., Williams, M., Heyeres, M., Lui, S. M. C., Klieve, H., Cadet-James, Y., Baird, L., et al. (2019). Assessing research impact: Australian research council criteria and the case of family wellbeing research. *Evaluation and program planning*, 73:176–186.

Tsey, K. (2019). Planning for and tracking research impact: Australian research council framework. In *Working on Wicked Problems*, pages 65–74. Springer.

Van Raan, A. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36(3):397–420.

Van Raan, A. F. (2004). Measuring science. In *Handbook of quantitative science and technology research*, pages 19–50. Springer.

Vanclay, F. (2003). International principles for social impact assessment. *Impact assessment and project appraisal*, 21(1):5–12.

Vanclay, F. (2006). Principles for social impact assessment: A critical comparison between the international and us documents. *Environmental Impact Assessment Review*, 26(1):3–14.

Witt, A., Diesner, J., Steffen, D., Rezapour, R., Bopp, J., Fiedler, N., Köller, C., Raster, M., and Wockenfuß, J. (2018). Impact of scientific research beyond academia: an alternative classification schema. *Proceedings of the LREC 2018 Workshop on Computational Impact Detection from Text Data*, pages 34–39.

Wolf, B., Lindenthal, T., Szerencsits, M., Holbrook, J. B., and Heß, J. (2013). Evaluating research beyond scientific impacthow to include criteria for productive interactions and impact on practice and society. *GAIA-Ecological Perspectives for Science and Society*, 22(2):104–114.

Wyndham, J. M. and Vitullo, M. W. (2018). Define the human right to science. *Science*, 362(6418):975–975.