Originally published in: Calzolari, Nicoletta/Kan, Min-Yen/Hoste, Veronique/Lenci, Alessandro/Sakti, Sakriani/Xue, Nianwen (Eds.): The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Paris: European Language Resources Association (ELRA), 2024. Pp. 4744-4749.

Detecting Impact Relevant Sections in Scientific Research

Maria Becker[1], Kanyao Han[2], Antonina Werthmann[1], Rezvaneh Rezapour[3], Haejin Lee[2], Jana Diesner[4], and Andreas Witt[1]

[1]Leibniz Institute for the German Language, Department of Digital Linguistics
[2]School of Information Sciences, University of Illinois Urbana Champaign
[3]Department of Information Science, Drexel University
[4]Technical University of Munich
(maria.becker|werthmann|witt)@ids-mannheim.de, (kanyaoh2|haejin2|)@illinois.edu,
shadi.rezapour@drexel.edu, jana.diesner@tum.de

Abstract

Impact assessment is an evolving area of research that aims at measuring and predicting the potential effects of projects or programs on a variety of stakeholders. While measuring the impact of scientific research is a vibrant subdomain of impact assessment, a recurring obstacle in this specific area is the lack of an efficient framework that facilitates labeling and analysis of lengthy reports. To address this issue, we propose, implement, and evaluate a framework for automatically assessing the impact of scientific research projects by identifying pertinent sections in research reports that indicate potential impact. We leverage a mixed-method approach that combines manual annotation with supervised machine learning to extract these passages from project reports. We experiment with different machine learning algorithms, including traditional statistical models as well as pre-trained transformer language models. Our results show that our proposed method achieves accuracy scores up to 0.81, and that our method is generalizable to scientific research from different domains and different languages.

Keywords: impact detection, project reports, annotation, mixed-methods, machine learning

1. Introduction

Scientific research can impact society and people in many ways, such as the development or adjustment of policies, practices, behavior, attitudes, and health. This collective influence is referred to as the social impact of research (Bornmann, 2013, 2012). The assessment of social impact has become increasingly crucial as funding agencies and governmental bodies demand evidence of the societal value of their investments in research (Williams and Grant, 2018; Heyeres et al., 2019; Gomes and Stavropoulou, 2019; Diesner et al., 2014; Diesner and Rezapour, 2015).

Impact assessment is an evolving, cross-disciplinary area of research and practice that aims at operationalizing, measuring, and predicting the potential positive and negative effects of a project, policy, or program on society, the environment, politics, or the economy among others (Bornmann and Daniel, 2005; Becker, 2001; Vanclay, 2006; Rezapour and Diesner, 2017). Measuring the impact of scientific research is a vibrant subfield of impact assessment, which, until now, mainly relied on analyzing research publications and their dissemination (e.g., citation counts) (Wildgaard et al., 2014; Wouters and Costas, 2012). This traditional approach, while valuable, often overlooks the multifaceted implications that research can have on society, policy, and industry. To address this gap, one area of research has been studying biases in publication trends (Way et al., 2019; Mishra et al., 2018) and the impact of research design choices and error propagation on findings about publishing behavior ((Kim and Diesner, 2017; Kim et al., 2014)). In order to capture and measure additional and broader types of impacts of science on society, Witt et al. (2018) proposed an alternative classification schema that aims to evaluate the impact of research projects beyond academia. This schema includes categories such as financial, technical impact, and environmental impact. Instead of mining scientific publications, they leverage project reports that are rarely used or shared outside of academia. Building on this work, Rezapour et al. (2020) analyzed a corpus of project reports and applied supervised machine learning to infer various impact categories from project reports. Their results show that various types of impacts are predictable from unseen data, and impact perception differs depending on how they were derived.

An important point highlighted in earlier studies is the laborious nature of manually annotating research reports. This challenge is exacerbated by the extensive length and details of the documents. Annotators must go through the entire document to identify segments that pertain to impact, and then assign an appropriate impact category to each identified segment. This presents a significant challenge, as manual annotation requires substantial time and resources. To address this issue, in this paper, we propose a method for automatically de-

LREC-COLING 2024, pages 4744–4749

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-127100

tecting passages in project reports that indicate the potential impact of scientific research projects. We leverage a mixed-methods approach in which we manually label project reports for impact-relevant passages, use human-generated heuristics to preprocess texts, and apply supervised machine learning techniques to extract these passages from project reports. We experiment with different machine learning algorithms and optimize our methods with a list of impact-indicating keywords and heuristics. We show in our experiments that our method is generalizable to scientific research from different domains such as artificial intelligence, mobility, linguistics, and musicology, and that our method is robust across two different languages (German and English). Another advantage of our method is that it performs well with both state-of-the-art neural models as well as traditional statistical models (Random Forest Model), and therefore can deal with a small amount of training data.

Our work makes the following contributions: We propose and evaluate a model that detects and extracts the potential impact of scientific research from research reports. This model is language agnostic and works with little, cross-domain annotated data. Using our method, researchers can (i) identify impact-relevant passages from research projects, and (ii) utilize the extracted passages for detailed analysis such as manual or corpus-linguistic investigations, or for annotations of more fine-grained impact categories.

2. Related Work

The assessment of social impact can be complex and multifaceted, necessitating the use of both qualitative and quantitative methods: In academia, bibliometric metrics such as citation numbers and the h-index have been widely used to evaluate the impact and quality of research (Bornmann and Daniel, 2008). Alternative metrics, or altmetrics, which consider the dissemination and use of research outputs by mentions of them beyond the academic community, have also been increasingly used in the recent decade (Priem and Hemminger, 2010). Recent studies often use well-curated and/or well-structured bibliometric data (e.g., PubMed, Scopus, and Web of Science) (Singh et al., 2021) or social media data (e.g., X (formerly Twitter) data with meta-information about numbers of likes and retweets) (Ortega, 2017).

Moreover, delving into textual data produced by researchers, like academic articles and grant reports, enriches the process of impact evaluation. Such materials provide insights into both the implemented and anticipated impacts of research from the viewpoint of the researchers. These documents reveal not just the final outcomes, but also the evolving trajectory and orientation of research efforts (Rezapour et al., 2020; Witt et al., 2018). However and by virtue of their very nature, academic articles and grant reports are often lengthy and may contain plenty of information irrelevant to impact assessment, including but not limited to project budgets, bibliographies, and logistic communication (Han et al., 2023). Pre-processing documents and extracting impact-relevant parts are hence crucial for further impact annotation and analysis. In view of this, our paper aims to advances this field by providing a method and pipeline for project report pre-processing and impact-relevant passage extraction.

3. Experiments

The task of detecting impact-relevant passages can formally be described as a binary classification task: the objective is to determine whether a given text segment signifies any form of impact or not.

3.1. Data

When a research grant concludes, the corresponding results are often consolidated in a project report and digitally archived. Unlike scientific papers, these reports are infrequently utilized as sources of scientific study and knowledge, therefore making new ways of automated processing especially beneficial. These reports, which typically range from 50 to 150 pages, detail the objectives, methodologies, findings, and conclusions. Additionally, the report explains the rationale behind the research and its potential future applications.

In this study, we utilize such reports for the analysis of impact. Our corpus consists of 1,160 German research reports from the fields of AI, automobility, linguistics, and musicology, accessible through the Leibniz Information Centre for Science and Technology (TIB). To use the reports, we first processed the files by converting the PDF documents to txt-files. We manually examined a sample of the converted files and identified special tokens that typically signal transitions between paragraphs, such as newlines and page numbers. Whenever we detected a special token, we divided the text accordingly. In some cases, a single paragraph might be split into two segments due to the presence of a page break or a table/graph. Each segment was considered an individual passage and was subject to separate annotation and automated detection.

3.2. Annotation

To produce labeled training data for our model, we selected a subset of 20 reports, with five reports from each domain, totaling 4,428 passages. While the number of selected reports is low, the length of the reports still leads to a sizeable datset. All passages were manually annotated by annotators with a linguistics background through a three-round

Domain	Example
AI	*Within this framework, the additional description of interfaces and integration possibilities in upstream or downstream value-added processes and chains ensures that not only the industries and use cases represented in the project benefit but also other industries.*
Automobility	*Another outcome of the project was the derivation of recommendations for safety standards for electric vehicles based on the results of user studies.*
Linguistics	*It can be stated that the didactic concept developed for the intervention study proves to be effective in helping students develop their pragmatic competencies, skills, and abilities.*
Musicology	*During the project a hybrid search and recommendation system was created. This allows the use of automated annotations and information manually annotated by the user to improve music search and recommendation.*

Table 1: Examples of (extracts from) impact relevant passages from our dataset.

annotation process: In the first round, Annotator 1 marked all passages that indicated impact in the document. Then Annotator 2 cross-checked the marked documents and corrected the annotations, if necessary. In the third round, Annotator 3 had a final look at the documents in order to harmonize and finalize the annotations.[1] Our final set of data consists of 1,661 passages annotated as impact-relevant and 2,767 labeled as non-impact-relevant. Examples of impact-relevant passages are given in Table 1. To test whether this method is robust to other languages, we created an English dataset by translating the German texts into English. We have checked the quality of the translations in detail on a subset of documents. We tried different translation tools, compared their performance in a pilot study, and then chose DeepL[2] as it had produced the best performing translation. Finally, we transferred the labels from the German texts to the translated dataset.[3]

3.3. Data Pre-Processing

We manually checked the passages that had been split automatically and found that some of them were irrelevant for impact analysis. Such passages were, for example, tables of contents (ToCs), bibliographies, headings, captions, and lists. To identify and exclude such passages automatically, we developed heuristics, including the detection of ToCs and bibliographies, which we remove together with the passages that appear before (ToCs) and after (bibliography) them. Additionally, we exclude all passages that contain less than 25 tokens (since they usually contain fragments, headings, or captions). This pre-processing step is not only helpful for removing passages that are irrelevant to the analysis of impact but also leads to cleaner data for model training. This step resulted in 2,498 passages for model building and testing. The mean and median numbers of words per passage in the final data were 96.5 and 87, respectively. We then built our models on German and English datasets respectively, and for each dataset, split the data into a training set (2054 passages – including 25% validation set) and a testing set (444 passages).

3.4. Models and Setup

We conduct experiments with both traditional classification model training, such as Random Forest (RF), Naive Bayes, and Logistic Regression, as well as model fine-tuning based on state-of-the-art pre-trained models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), which are trained on a huge amount of data (BookCorpus, Wikipedia, CC-News, among others) and have hundreds of millions of parameters.

Traditional classification model training. As features for the traditional classification models, we create a list of words that signal impact that we identified as follows: using the words *Einfluss* (*influence*) and *Impact* (*impact*) as seeds, we employed the German co-occurrence database CCDB[4] to find words with similar co-occurrence profiles (and therefore are likely to have a similar meaning). We manually checked the results and created a list of 93 words, which we then automatically translated to English in DeepL. This list contained words such as *effect, progress, improve, efficient* and *success*. The complete lists of German and English words are provided in the Appendix. We then calculated the frequency of each signal word in each document via keyword search , and usde them as features. To achieve the best performance, in addition to impact-related words, we also experimented with different hyperparameters, including the maximum depth of the tree, minimum number of samples required to split an internal node, penalty methods and strengths, and class weights.

Transformer model fine-tuning. For the English dataset, we used the Cased English BERT base

[1] Since the documents were not annotated simultaneously but rather in a multi-step annotation procedure, we can't report Inter Annotator Agreement (IAA) for this task.

[2] https://www.deepl.com/translator

[3] Another option would have been to annotate similar texts in English that don't need to be translated. Since we are not aware of existing corpora that are comparable, for our approach, we relied on the automatic translation and the transfer of labels.

[4] http://corpora.ids-mannheim.de/ccdb/

Parameters of the Random Forest Models			
n_estimators	1000	min_samples_leaf	6
class_weight	bal.	max_depth	6

Parameters of the Transformer Models			
num_train_epochs	10	weight_decay	0.01
train_batch_size	16	logging_steps	100
eval_batch_size	16	save_steps	100
warmup_ratio	0.1	learning_rate	1e-5

Table 2: Model hyperparameters.

model and its advanced version, the RoBERTa base model, as pre-trained models. However, as there is no German RoBERTa model available, we only fine-tuned the Cased German BERT base model for our German dataset. To ensure comparability between the English and German datasets, we adopted the same hyperparameters for both models, including batch size, learning rate, and logging steps. We experimented with different hyperparameters and report the best hyperparameters in Table 2.

4. Evaluation, Results and Discussion

4.1. Quantitative Evaluation

We assessed model performance by precision, recall, F1, and accuracy scores based on the test set. For the traditional statistical models, we only show the performance from the best classifier, i.e. Random Forest (RF). Results are shown in Table 3.

Comparing the performance of **RF models** on the German versus English datasets shows that the results are comparable. The slightly lower scores of RF English could possibly be traced back to the automatic translation process, which can introduce errors that can lower the quality of the training/testing data.

In contrast to RF German versus RF English, **BERT** English performed slightly better than BERT German. This might be attributable to the fact that the performance of fine-tuned models heavily relies on the amount of training data (BERT English is pretrained on a larger dataset than BERT German). Furthermore, the overall performance of the fine-tuned **RoBERTa** model (which is trained on ten times more data than BERT) is the best among all models, though only available for English.

The **comparison between traditional models and large language models** shows mixed results. We find that for our data, large language models do not necessarily outperform the traditional statistical RF model. Specifically, the RF model has a higher Recall than Precision, while the transformer models like BERT and RoBERTa show the opposite trend. This indicates that while transformer models have fewer false positives, they might overlook some impact-relevant passages. On the other hand, the RF model identifies more of these passages but also produces more false positives along the way. Our recommendation for researchers using our models is to select a model based on the research objectives: if aiming for detailed annotation during which annotators could further filter out false positives, the RF model, with its broader coverage, is preferable. However, for creating a dataset that solely contains impact-relevant passages, transformer models offer better precision.

Finally, we investigated the **performance of the RF model** (in terms of Class 1) **in various domains**. For the domains of AI and Automobility, the RF model showed higher recall (0.72-0.73) than the RF model for Linguistics and Musicology (0.61-0.62) when tested on German data. This recall difference was also observed with the German BERT model. These results imply that there is a higher likelihood of missing impact-relevant passages in the humanities and social sciences than in technology and engineering fields.

4.2. Manual error analysis

To identify impact-indicating instances that our models failed to detect, we conducted a manual evaluation of 30 randomly selected false negatives spanning all four domains for each model (RF and BERT). For each passage, we (i) made an assumption about why the model could have overseen the impact by looking at the passage to find out (linguistic) peculiarities that could have an effect on the classification task, and (ii) investigated the type of impact mentioned in each passage out of seven categories: societal, economical, environmental, ethical, legal, technical, and academic impact; plus the residual category *other impact*.

We found that many of the passages misclassified by the **RF model** were incomplete sentences (e.g., due to segmentation errors or bullet-point-style formulations; 26% of the 30 evaluated passages). In 19% of the passages, relevant context was missing, and 26 % contained unknown words (abbreviations or spellings). The most frequent characteristics in the passages misclassified by **BERT** were also incomplete sentences (31%) and missing contextual information (31%), followed by unknown words (17%). Also, 10% of the passages that BERT misclassified contain citations in English, which could also be interpreted as an error source.

Regarding the actual **impact categories**, we found that both models struggled with passages that matched none of the seven categories listed above: 39% of the 30 passages misclassified by the RF model and 50% of the 30 passages misclassified by BERT belonged to the residual category *other impact*. This finding further supports the suitability of our previously developed impact categorisation model. Moreover, passages that expressed tech-

	Precision	Recall	F1	Accuracy	SE
RF German	0.41 (0.63)	0.66 (0.67)	0.51 (0.63)	0.67	0.039
RF English	0.40 (0.63)	0.64 (0.66)	0.49 (0.63)	0.67	0.044
BERT German	0.62 (0.72)	0.38 (0.65)	0.47 (0.67)	0.78	0.038
BERT English	0.63 (0.73)	0.42 (0.67)	0.50 (0.69)	0.80	0.037
RoBERTa English	0.59 (0.74)	0.66 (0.76)	0.62 (0.75)	0.81	0.036

Table 3: Model performance for Class 1 (impact relevant passages). The macro average scores for Class 1 + Class 0 (passages that are not impact relevant) are presented in parentheses. SE is the standard error of accuracy with a 95% confidence interval.

nical impact were often not recognized as being impact relevant by the RF model (44 % of all 30 passages), while BERT often misses passages that mention academic impact (27%).

5. Conclusion

In this paper, we proposed a mixed-methods approach for automatically assessing the impact of scientific research projects by identifying sections in project reports that indicate potential impact of the work on XXX. Our dataset comprises reports from four different research domains in both German and English. We split these reports into passages that we manually annotated for whether they expressed impact of a given project or not. We then used the labeled data for training models with a supervised machine learning approach. We experimented with both traditional classification models, which we enhanced with signal words as features, and pretrained transformer language models, which we fine-tuned for our task. Our experimental results showed that our proposed method is generalizable to scientific research from different domains as well as from different languages, and works with little annotated data. Using our method, researchers can assess the overall impact potential of their research projects, and utilize the extracted passages for further analysis.[5]

6. Limitations

Our research has several limitations. First, the scope of our dataset is restricted to a few fields and domains and therefore lacks topical diversity. While our impact definition aims to be broadly applicable, its relevance is specifically evaluated within these narrow domains, suggesting the necessity for broader generalization in future studies. Additionally, our English dataset was constructed using an off-the-shelf translation tool. Despite recognizing the potential drawbacks of this method, we opted for this solution due to the absence of directly comparable data. This decision allowed us to mirror the unique nature and context of our original German dataset. Furthermore, our study uses a limited array of classification models for identifying impact-relevant passages within the data. While incorporating cutting-edge generative AI models, such as those in the GPT series, might improve accuracy, we consciously opted for simplicity and practicality over innovation, thus limiting ourselves to the models explicitly described in our study. This decision acknowledges the trade-off between the potential for improved results and the introduction of complexities that are beyond the scope of this study.

[5]The dataset, code, and model are available at: https://doi.org/10.13012/B2IDB-9934303_V1

7. Acknowledgements

We would like to thank Jason Brockmeyer for his contribution to and help with this project. This work is part of the project “TextTransfer Corpus based detection of secondary usage of scientific publications”, which is funded 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 thank the annotators for their tremendous help and input throughout this project.

8. Bibliographical References

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

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

Lutz Bornmann. 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.

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

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Jana Diesner and Rezvaneh Rezapour. 2015. Social computing for impact assessment of social change projects. In *Social Computing, Behavioral-Cultural Modeling, and Prediction: 8th International Conference, SBP 2015, Washington, DC, USA, March 31-April 3, 2015. Proceedings 8*, pages 34–43. Springer.

Daniela Gomes and Charitini Stavropoulou. 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.

Kanyao Han, Rezvaneh Rezapour, Katia Nakamura, Dikshya Devkota, Daniel C Miller, and Jana Diesner. 2023. An expert-in-the-loop method for domain-specific document categorization based on small training data. *Journal of the Association for Information Science and Technology*, 74(6):669–684.

Marion Heyeres, Komla Tsey, Yinghong Yang, Li Yan, and Hua Jiang. 2019. The characteristics and reporting quality of research impact case studies: A systematic review. *Evaluation and program planning*, 73:10–23.

Jinseok Kim and Jana Diesner. 2017. Over-time measurement of triadic closure in coauthorship networks. *Social Network Analysis and Mining*, 7:1–12.

Jinseok Kim, Heejun Kim, and Jana Diesner. 2014. The impact of name ambiguity on properties of coauthorship networks. *Journal of Information Science Theory and Practice*, 2(2):6–15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shubhanshu Mishra, Brent D Fegley, Jana Diesner, and Vetle I Torvik. 2018. Self-citation is the hallmark of productive authors, of any gender. *PloS one*, 13(9):e0195773.

Jose Luis Ortega. 2017. The presence of academic journals on twitter and its relationship with dissemination (tweets) and research impact (citations). *Aslib journal of information management*, 69(6):674–687.

Jason Priem and Bradely H Hemminger. 2010. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First monday*.

Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Diana Steffen, Andreas Witt, and Jana Diesner. 2020. Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6777–6785, Marseille, France. European Language Resources Association.

Rezvaneh Rezapour and Jana Diesner. 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.

Vivek Kumar Singh, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. 2021. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126:5113–5142.

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

Samuel F Way, Allison C Morgan, Daniel B Larremore, and Aaron Clauset. 2019. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, 116(22):10729–10733.

Lorna Wildgaard, Jesper W Schneider, and Birger Larsen. 2014. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, 101:125–158.

Kate Williams and Jonathan Grant. 2018. A comparative review of how the policy and procedures to assess research impact evolved in australia and the uk. *Research Evaluation*, 27(2):93–105.

Andreas Witt, Jana Diesner, Diana Steffen, Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler, Christoph Köller, Manu Raster, and Jennifer Wockenfuß. 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.

Paul Wouters and Rodrigo Costas. 2012. Users, narcissism and control: tracking the impact of scholarly publications in the 21st century.