LREC 2018 Workshop

1st Workshop on Computational Impact Detection from Text Data

PROCEEDINGS

Edited by

Jana Diesner, Georg Rehm, Andreas Witt

ISBN: 979-10-95546-05-4
EAN: 9791095546054

08 May 2018

Proceedings of the LREC 2018 Workshop
1st Workshop on Computational Impact Detection from Text Data

08 May 2018 – Miyazaki, Japan

Edited by Jana Diesner, Georg Rehm and Andreas Witt

Organising Committee

- Jana Diesner, The iSchool at University of Illinois at Urbana-Champaign, USA
- Georg Rehm, DFKI GmbH, Germany
- Andreas Witt, University of Cologne / IDS Mannheim / Heidelberg University, Germany

Programme Committee

- Jana Diesner, The iSchool at University of Illinois at Urbana-Champaign, USA
- Lindsay Green-Barber, The Impact Architects, San Francisco, USA
- Franciska de Jong, CLARIN ERIC, The Netherlands
- Sandra Kübler, University of Indiana Bloomington, USA
- Georg Rehm, DFKI GmbH, Germany
- Andreas Witt, University of Cologne / IDS Mannheim / Heidelberg University, Germany
- Feiyu Xu, Lenovo Institute of Artificial Intelligence Laboratory Research, China

Preface

How can we measure the impact – such as awareness for economic, ecological, and political matters – of information, such as scientific publications, user-generated content, and reports from the public administration, based on text data? This workshop brings together research from different theoretical paradigms and methodologies for the extraction of impact-relevant indicators from natural language text data and related meta-data. The papers in this workshop represent different types of expertise in different methods for analyzing text data; spanning the whole spectrum of qualitative, quantitative, and mixed methods techniques, as well as domain expertise in the field of impact measurement. The program was built to create an interdisciplinary half-day workshop where we discuss possibilities, limitations, and synergistic effects of different approaches.

These proceedings comprise the papers presented at the "First Workshop on Computational Impact Detection from Text Data", which is collocated with the "International Conference on Language Resources and Evaluation" (LREC 2018), which takes place in May 2018 in Miyazaki, Japan.

We are thankful to the authors who submitted to this workshop, our Program Committee members for their contributions, and LREC for including this workshop into their program.

J. Diesner, G. Rehm and A. Witt May 2018

Programme

09:00 – 09:05 Welcome and Introduction

Session 1

09:05 – 09:35 Zhongsheng Wang, Kiyoaki Shirai:
The Financial Attention Index to Measure Impact of Crisis from Microblog

09:35 – 10:05 Philipp Heinrich, Christoph Adrian, Olena Kalashnikova,
Fabian Schäfer, Stefan Evert:
A Transnational Analysis of News and Tweets about
Nuclear Phase-Out in the Aftermath of the Fukushima Incident

10:05 – 10:35 Drahomira Herrmannova, Petr Knoth,
Christopher Stahl, Robert Patton, Jack Wells:
Text and Graph Based Approach for Analyzing Patterns of
Research Collaboration: An analysis of the TrueImpactDataset

10:35 – 11:00 *Coffee Break*

Session 2

11:00 – 11:30 Zygmunt Vetulani, Marta Witkowska, Umut Canbolat:
TSCC: a New Tool to Create Lexically Saturated Text Subcorpora

11:30 – 12:00 Andreas Witt, Jana Diesner, Diana Steffen,
Rezvaneh Rezapour, Jutta Bopp, Norman Fiedler,
Christoph Köller, Manu Raster, Jennifer Wockenfuß:
Impact of Scientific Research beyond Academia:
An Alternative Classification Schema

12:00 – 12:30 Yue Chen, Kenneth Steimel, Everett Green, Nils Hjortnaes,
Zuoyu Tian, Daniel Dakota, Sandra Kübler:
Towards Determining Textual Characteristics of High and Low Impact Publications

12:30 – 13:00 Discussion and conclusions

Table of Contents

Towards Determining Textual Characteristics of High and Low Impact Publications

Yue Chen, Kenneth Steimel, Everett Green, Nils Hjortnaes, Zuoyu Tian, Daniel Dakota, Sandra Kübler
Indiana University
Bloomington, IN, USA
{yc59,ksteimel,evegreen,nhjortn,zuoytian,ddakota,skuebler}@indiana.edu

Abstract

This paper is concerned with the question of whether we can predict the future impact of a paper based on the text of the paper. We create a corpus of papers in computational linguistics, and we create gold standard impact annotations by using their Google Scholar citation counts. We use supervised classification approaches to automatically predict impact of the papers. Our results when using very simple features show some success, but they also show that the classifiers suffer from class imbalance problems.

Keywords: impact detection, data imbalance, corpus

1. Introduction

This paper is concerned with the question whether we can predict the future impact of a paper based on the text of the paper. In other words, are there textual characteristics that increase the impact of a paper? We define the impact of a paper as its citation count. While this question sounds somewhat unrealistic, it does make sense when looked at from the angle that properly advertising one's work should have a positive effect on its reception. A well written paper cannot succeed if there is no academic content. But some papers that have the content, but package it suboptimally may not get as much attention as they deserve. In this vein, our question can be rephrased as: Which textual characteristics do we need to adapt in order to produce a successful paper?

In our work, we investigate papers from the major conferences and journals in computational linguistics. We create a corpus of such papers on the topics of parsing and machine translation, and we create a gold standard of their impact by using their Google Scholar citation counts. We then separate the papers into three classes: low impact, high impact, and highest impact. We use supervised classification approaches to automatically predict impact of the papers. Our results when using very simple features show some success, especially when we use the full papers rather than just the abstracts, but they also show that the classifiers suffer from problems; they have a tendency to group all papers into the low impact class, which is the majority class.

There are two possible reasons for the behavior of the classifiers: One possibility is that the features we use are not predictive enough. The second possibility concerns the problem of class imbalance since the highest impact setting has very few examples. Depending on which of the reasons holds, we need to address the problem by either feature engineering or data sampling. To test the two hypotheses, We removed stopwords from the content, both abstracts and whole texts. We also experimented with both down-sampling and up-sampling. Random down-sampling of the low and high citation classes yields more balanced performance across the classes but results in a reduced overall accuracy due to the small amount of data used. This discrepancy is even more pronounced when only abstracts are used. Synthetic minority up-sampling techniques produced results very similar to the previous experiments.

The paper is structured as follows: We present related work in section 2., followed by a description of the corpus in section 3.. Section 4. presents the experiments and results with section 4.5. presenting an analysis of the features. We conclude with areas of future research in section 5..

2. Related Work

Traditional methods to determine the impact of a publication have heavily focused on citation counts. However, there are many methodological issues to consider as well as many caveats in these results. Furthermore, such metrics are often only retroactively obtainable and cannot indicate future impact. This has led to more focused work examining whether different sorts of features can be utilized to gauge the future impact of a publication. We are aware that this limits the objectivity of our gold standard (see section 3.), but since we are interested in automatic approaches to predicting future impact based on text, we assume that a switch in determining the gold standard will not have impact the usability of our methodology.

2.1. Citation Count Impact

Rankings based on citation counts are often used to demonstrate the "importance" of a publication. This is often performed by simply counting the number of times a publication (or a group of citations) has been cited by a different set of publications. More complex measures aim to account for types of variation and instead focus on the average number of citations on a set of papers and compensate for the length of time publication has been in existence (i.e. weighting publications having existed for three years against those for fifty years). Such methodology does yield a plethora of information. Adams et al. (2005) use citation probability metrics on the the Institute for Scientific Information to discover certain trends including: Higher ranked universities' citation sharing, mutual cross-over between scientific fields

in citations, and that there is a lag of about three years for the diffusion of information.

However, although informative and easy to access in terms of information, relying strictly on citation counts and probability metrics is often misleading and prone to inherent bias based on the given criteria. Meho and Yang (2007) compile a corpus created by fusing different citation metric systems, such as WoS and Scopus, to demonstrate that a selected metric significantly impacts how a publication can be ranked based on citation counts as different metrics exclude different fields, languages, or publication types.

Another approach taken is correlating the number of citations with the impact factor of the journal of publication to examine the interaction of the two. Levitt and Thelwall (2011) noted that standard citation metrics are not necessarily the best indicator of impact for the subject of economics, as there is also a strong correlation with the journal of publication. This suggests that the forum of publication is also relevant to impact, not just the number of citations and substance of the article.

2.2. Predictive Impact

Traditional methods of impact assessment can only be performed after a reasonable amount of time has passed to allow for the dissemination of the publication into a research community. Much of this work focuses on the use of citation counts to determine impact; however, this is rather limited in terms of future predictability. Thus, approaches utilizing more content for impact prediction have been an area of more recent research.

Ibáñez et al. (2009) examined which types of classifiers and features can be used to predict future citation frequency. They found that certain classifiers, such as Naive Bayes, performed better but also that certain tokens can actually be indicative of a publications of future citation frequency. Dietz et al. (2007) use an LDA-based approach that attempts to detect topical influence of cited documents on the citing document by linking individual references and word distributions on citing papers.

Other recent work has looked at how citation impact can be predicted at a publication's release. This has become relevant due to the electronic publication of many articles upon release. Brody et al. (2006) found a correlation between downloads of arXiv articles in certain scientific fields and their citation and impact. They further argue that downloads can also show a usage impact that is not correlated to citations and that as more databases become available, such impact may only increase.

With the advent of social media, the announcement of the existence of new publications is disseminated through these mediums. This was explored by Eysenbach (2011) who noted that Twitter can help predict high impact publications by the frequency a publication is tweeted within the first few days of publication, suggesting that non-traditional metrics can be used to immediately identify impact.

3. Impact Corpus

We are interested in whether the content of a paper can give us information on whether this paper will have an impact on the field. Since we did not find any corpus that would allow us to investigate this question, we created our corpus. The corpus was sampled from leading publications in the field of Computational Linguistics, and more specifically from major conferences and journals that are incorporated into the ACL Anthology[1]. Specifically, we only took papers from the Computational Linguistics journal, ACL, NAACL, EACL and EMNLP due to their content and stylistic similarities. Since we need to access the text, using the PDFs from the anthology directly is of limited use. Thus, we used the texts available from the ACL Anthology Network[2] for the textual basis. This corpus was created by using OCR to convert the PDFs into text, with additional post-processing using both scripts and manual labor (Radev et al., 2009; Radev et al., 2013). We decided to concentrate on two major topics of computational linguistics, parsing and machine translation. To extract papers on those topics, all texts that use the words "parse" or "parsing" (case invariant) in their title were extracted for the parsing category, and all papers using the words "translate", or "translation" in the title were extracted for the machine translation category[3].

Year	Total Papers	Parsing	Machine Translation
2007	187	83	104
2008	279	108	171
2009	270	135	135
2010	306	130	176
2011	191	67	124
2012	225	81	144

Table 1: Distribution of papers across years

Since we define the impact of paper in terms of the number of citations a paper has received, we need to allow sufficient time between publication of the original paper and of the papers citing it. Thus, we chose a window of 5 to 10 years ago, i.e., we consider papers published between 2007 and 2012. Table 1 displays the distribution of papers across the years for which we collected data.

Citation counts were then collected for each of these papers using Google Scholar[4]. We extracted the citation counts manually, and we list the sum of all citations if a paper is listed more than once on Google Scholar.

Figure 1 shows the distribution of citation counts in the two topics. Based on this distribution, we established three categories: a low citation count (0-29 citations), a high citation count (30-119), and an extremely high citation count (>120). The graphs in Figure 1 show that this split results in a severely imbalanced data set, which will make the automatic prediction of impact very challenging. Citation counts follow a rough Zipfian distribution: 948 papers fall into the low-count bin, 424 papers fall into the high-count

[1] http://aclweb.org/anthology/

[2] http://tangra.cs.yale.edu/newaan/

[3] A third category corresponding to stance detection/sentiment analysis was also collected. However, the resulting collection of papers was too small to be of use.

[4] https://scholar.google.com/

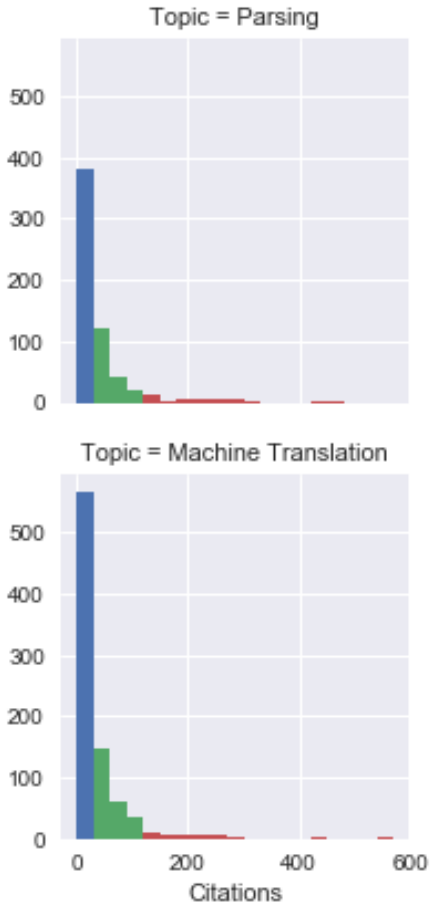


Figure 1: Citation class distribution with 3 classes

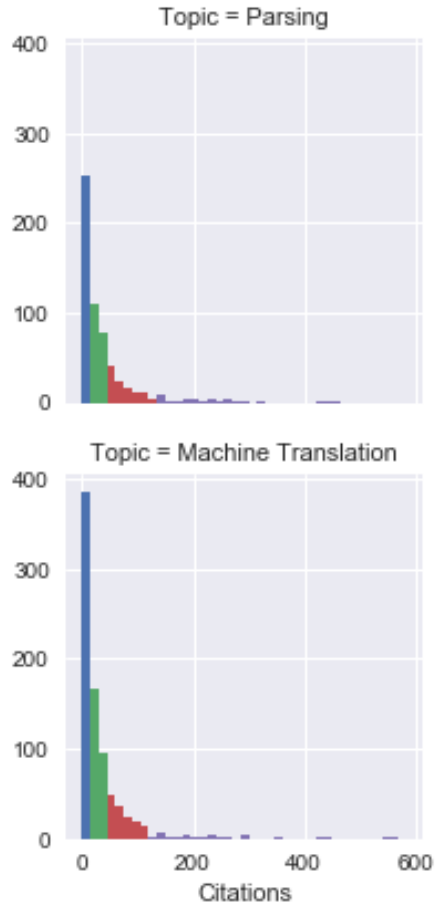


Figure 2: The 5 class split

Topic	Low	High	Highest
Parsing	381	181	42
MT	567	243	44

Table 2: Class distribution by topic

bin and 86 papers fall into the final highest-count bin. Table 2 shows the distribution across the two topics.

We are looking into an alternative classification using five citation classes as a way to mitigate the imbalance in the data. The classes consist of a no impact class for papers receiving 0 citations, a low class for papers with 1-15 citations, a moderate class for papers with 16-45, a high class for papers with 46-125 and a very high class for papers with more than 125 citations. The number of papers for this 5-class system are shown in figure 2. This graph shows that we obtain a less skewed data set.

4. Experiments

Our interest is whether we can predict a paper's future impact based on characteristics in the paper. We conducted a series of experiments to investigate how well the impact class can be predicted based on characteristics of the text in papers. For these experiments, we use a simple bag of words approach. All of the experiments presented here are based on the skewed 3-class data split.

We experiment with two types of texts: paper abstracts and full texts. This will ultimately answer the questions whether we can determine the impact of a paper based solely on the abstract and whether the abstract is as informative as the full paper. We also experiment with an additional pre-processing step: removing the stopwords. The list of stopwords is obtained from NLTK (Bird et al., 2009).

4.1. Extracting Abstracts

We extract abstracts automatically from the corpus using regular expressions. The regular expression will take the texts between the word "Abstract" and the word "Introduction". As some of the papers do not follow this pattern, their abstracts were not extracted. For these 70 papers, the abstracts could not be identified successfully, therefore we extracted those abstracts manually.

4.2. Experimental Setup

To create the training, development and test datasets used, we split the corpus per topic, i.e., we created separate training, development, and test sets for the parsing and the MT

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Random Forest	All	**60.61**	**45.74**	All	67.82	56.87
	10 000	**60.61**	**46.18**	3 000	68.97	59.15
Gradient Boost Trees	All	**60.60**	**46.18**	All	65.52	56.87
	5 000	60.60	48.91	10 000	67.82	58.89
Adaptive Boosting	All	**60.60**	**45.74**	All	66.67	58.89
	4 000	**60.60**	**45.74**	2 000	67.82	61.79
SVM	All	62.12	53.67	All	68.97	65.66
	10 000	62.12	57.22	10 000	68.97	65.66

Table 3: Results for both topics using only the papers' abstracts (boldface: majority classification)

domain. Out of every 10 papers, we randomly selected 1 paper for the development dataset, 1 paper for the test data, and the remaining 8 papers for the training set.

For the features, we extracted word unigram, bigram, and trigram counts from the texts of the training set. In the experiments shown in section 4.3., only the abstract is used for feature extraction while the experiments in section 4.4. use the entire text including the abstract. Then, we performed feature selection via a filter method, using both χ^2-goodness of fit and Mutual Information. The features with the highest scores below a specified count threshold are kept while all others are removed. We only report results using Mutual Information. χ^2 tends to result in similar, occasionally somewhat lower performance.

To gauge how sensitive performance is to specific machine learning approaches, we experiment with a variety of classification algorithms: Random Forest, Support Vector Machines (SVM), Adaptive Boosting, and Gradient Boost using shallow decision trees. We use the implementation in `scikit-learn` (Pedregosa et al., 2011). Each of the classifiers is trained using an exhaustive search over hyperparameter values.

4.3. Classifying Abstracts

The results for both topics are shown in table 3. The table shows several interesting results: First, it is clear from looking at accuracy that word n-grams do not provide enough information to determine impact of papers reliably. Additionally, the F-scores are considerably lower than the accuracies. This difference gives us an indication one problem: Many of the results are based on majority classification, i.e., the machine learner exclusively chooses the class that constitutes the majority class in the training data. Such cases are indicated in bold in the table. This shows that most of the classifiers prefer majority classification. Feature selection, which has been shown to have the potential of being useful in problems with class imbalance (Kübler et al., 2017), does not have any effect on accuracy in parsing. For the machine translation topic, it has a positive effect on all ensemble methods but does not improve the accuracy of SVMs. We will return to the question of majority classification below and have a closer look at performance per class.

As we described above, we repeated the experiments after having removed the stopwords. For abstract only features, this pre-processing step did not help with either accuracy or F-score.

A second trend that is obvious from table 3 is that predicting impact for the machine translation topic is more successful than for parsing: The highest accuracy reaches almost 70% while for parsing, the highest accuracy is around 62%. This cannot be explained by the imbalance in the data since machine translation has a higher skewing factor (the majority class is 1.98 times more likely than the other two classes combined) than parsing (1.71 times). Especially for SVMs, the F-scores are close to the accuracies, which means that the classifier goes beyond majority classification.

Returning to the issue of majority classification, table 4 shows the results in terms of precision and recall for selected experiments. These results show how serious the issue is: for parsing, SVM and Gradient Boosted Trees are the only classifiers that can identify at least some papers in the High class. For machine translation, all classifiers successfully identify some of the High class. However, none of the settings identifies any of the papers in the Highest class. At this point, it is unclear whether this is a consequence of the class imbalance in the data set or whether our feature set is not expressive enough to distinguish the classes. Further experiments using methods to address class imbalance are needed.

4.4. Classifying Full Papers

We now turn to the experiments where we use the full text instead of abstracts. The results of those experiments are shown in table 5. These results show that predicting impact based on the full text is more successful than predictions using only the abstract: For parsing, Adaptive Boosting reaches an accuracy of 77.27%, which is about 17% absolute higher than for abstracts. For machine translation, the same classifier reaches 72.41%, which is 5% absolute higher than its results on abstracts. Interestingly, both of these results are based on a small number of features chosen by feature selection. The corresponding F-scores show similar trends.

The results for the experiments in which we removed the stopwords are shown in table 6. We focus on the same settings as in table 5 to allow for a direct comparison between the two settings. These results show several interesting trends: For parsing, removing stopwords results in a massive deterioration across all classifiers. For machine translation, in contrast, Adaptive Boosting shows a minimal gain of 0.8% absolute in terms of F-score, and Random Forest gains close to 8% absolute. The reason for these gains

Topic	Classifier	# features	Class	Precision	Recall
Parsing	Random Forest	10 000	Low	61.54	100.00
			High	0.00	0.00
			Highest	0.00	0.00
	SVM	10 000	Low	67.31	87.50
			High	42.86	27.27
			Highest	0.00	0.00
	Gradient Boost Trees	5 000	Low	62.90	97.50
			High	25.00	4.55
			Highest	0.00	0.00
	Adaptive Boosting	4 000	Low	60.61	100
			High	0.00	0.00
			Highest	0.00	0.00
Machine translation	Random Forest	3 000	Low	67.86	100.00
			High	100.00	12.00
			Highest	0.00	0.00
	SVM	All	Low	71.83	89.47
			High	64.29	34.62
			Highest	0.00	0.00
	Gradient Boost Trees	10 000	Low	72.05	85.96
			High	62.50	38.46
			Highest	0.00	0.00
	Adaptive Boosting	2 000	Low	68.83	92.98
			High	60.00	23.08
			Highest	0.00	0.00

Table 4: Per class precision and recall for abstracts

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Adaptive Boosting	1 000	77.27	74.56	3 000	72.41	65.38
Support Vector Machines	50 000	68.18	60.35	50 000	71.26	64.43
Random Forest	2 000	71.21	65.56	1 000	71.26	64.50

Table 5: Results for both topics using the whole text (including stopwords)

	Parsing			Machine Translation		
Classifier	# features	Accuracy	F-score	# features	Accuracy	F-score
Adaptive Boosting	1 000	57.58	56.61	3 000	71.26	66.19
Support Vector Machines	50 000	54.55	56.36	50 000	59.77	58.73
Random Forest	2 000	66.67	60.00	1 000	74.71	72.44

Table 6: Results for both topics using the whole text (no stopwords)

require further investigation.

Table 7 shows the results in terms of precision and recall per class. These results corroborate our findings from table 5: The classifiers are all more successful in identifying High Impact papers than when they only have access to abstracts. This means that full papers contain more information about whether a paper has future impact on the field. When we allow stopwords in the features set, we do not find any of the Highest Impact papers. When we remove stopwords, however, SVM is able to predict the highest class with a precision of 12.50% and a recall of 25.00%. Even though these results are not stellar, we find this very encouraging in that feature engineering shows some impact on finding these highly cited papers.

4.5. Feature Analysis

Here we examine what types of features are selected by the feature selection model on abstracts. We focus on general trends within the features and potential correlations with known real world events during the selected time frame.
We first have a look at the experiments using abstracts only. For the top features for parsing, it is easier to identify common patterns and trends than for their machine translation counterparts. For example, the CoNLL 2007 shared task (Nivre et al., 2007) played an influential role in the direction of the field and is aligned with our time interval. This is noted in the returned features return for parsing as

				With Stopwords		Without Stopwords	
Topic	Classifier	# features	Class	Precision	Recall	Precision	Recall
Parsing	Adaptive Boosting	1 000	Low	79.17	95.00	70.00	70.00
			High	76.47	59.09	40.00	45.45
			Highest	0	0	0	0
	Support Vector Machines	50 000	Low	65.57	100.00	74.29	65.00
			High	100.00	22.73	39.13	40.90
			Highest	0	0	12.50	25.00
	Random Forest	2 000	Low	71.43	100.00	69.64	97.50
			High	77.78	31.82	33.33	9.09
			Highest	0	0	0	0
Machine Translation	Adaptive Boosting	3 000	Low	71.25	100.00	75.81	82.46
			High	85.71	23.08	52.00	50.00
			Highest	0	0	0	0
	Support Vector Machines	50 000	Low	70.00	100.00	70.49	75.44
			High	85.71	23.08	39.13	34.62
			Highest	0	0	0	0
	Random Forest	1 000	Low	70.89	98.25	78.13	87.72
			High	75.00	23.08	65.23	57.69
			Highest	0	0	0	0

Table 7: Precision and recall using the whole text

not only are references to the shared task returned, but many related terms: multilingual, dependency parsing, track. Not only was the shared task influential, but many of the then state-of-the-art systems participated in the task. This explains why so many of the top features can easily be associated with this knowledge. This leads to an interesting aspect: that by taking a small time interval, the currently most prominent topics will lead to the highest correlation to impact. One way to address this issue may be the use of topic modeling, for modeling this association between current topics of interest in the field and citation counts.This needs to be investigated further.

The features selected for machine translation, however, are not particularly informative. In the experiments using stopwords, many of the high-ranking features for MT are stopwords: the, we, to. While it is possible that certain grammatical constructions may be more clear, and thus papers that use these constructions may be cited more often, it does not seem likely. Comparing the features returned by both Mutual Information and Chi-square do not yield particularly interpretable features. In the experiments disregarding stopwords, many features can easily be associated with the field in general: system, evaluation, domain. Such features should provide any value in distinguishing between different levels of impact. This would help explain why there is little improvement gained without adding large quantities of features.

One exception is "Joshua" which refers to an MT system released in 2009 (Li et al., 2009) and is returned as a high ranking feature. This is interesting given that it was intended to be an alternative to the MT system "Moses" (Koehn et al., 2007) released in 2007 which is also a returned feature but with a much lower ranking. This further supports the notion of using topics as features for the classifier may give us access to current trends in the field as an indication of high impact features. However, one downside to using topics in this manner is that these topics may be too specific to a given time interval, and would not have the same usefulness in terms of determining impact for publication during a different era given that trends change.

5. Future Work

We have only scratched the surface of the problem of identifying the future impact of a paper based on textual features only. More experimentation and examination of the features is still required, particularly with regard to preprocessing decisions and the additions of various types of representations (e.g., lemmatization). We predict that these preprocessing decisions will have a strong impact on our results. Unlike many prediction tasks in which text is often shorter or limited (such as opinion mining of Twitter data), more text is available, thus there is a need to determine the best way of preprocessing such texts to eliminate as much noise as possible while also keeping specific types of non-standard features (e.g., keeping track of the number of figures or tables).

Additionally, while we see some success of classifiers in predicting high impact papers, we need to investigate whether other feature types are useful or whether we can improve results by using methods to address class imbalance in the data. Additional feature types will include character n-grams, which have been used successfully in stance detection tasks with imbalanced data (Mohammad et al., 2016), but also dependency triples and chains. Class imbalance can be addressed by upsampling methods that create artificial examples. In addition, the five way split described in section 3. may balance the classes better.

6. Bibliographical References

Adams, J. D., Clemmons, J. R., and Stephan, P. E. (2005). Standing on academic shoulders: Measuring scientific

influence in universities. *Annales d'Économie et de Statistique*, pages 61–90.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Brody, T., Harnad, S., and Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 233–240, Corvalis, OR.

Eysenbach, G. (2011). Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4):e123, December.

Ibáñez, A., Larrañaga, P., and Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, 25(24):3303–3309.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Kübler, S., Liu, C., and Sayyed, Z. A. (2017). To use or not to use: Feature selection for sentiment analysis of highly imbalanced data. *Natural Language Engineering*.

Levitt, J. M. and Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing and Management*, 47(2):300–308.

Li, Z., Callison-Burch, C., Dyer, C., Khundapur, S., Schwartz, L., Thornton, W., Weese, J., and Zaidan, O. (2009). Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece.

Meho, L. I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, CA.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL 2007 Shared Task. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. (2009). The ACL Anthology Network Corpus. In *Proceedings of the ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore.

Radev, D., Muthukrishnan, P., Qazvinian, V., and Abu-Jbara, A. (2013). The ACL Anthology Network Corpus. *Language Resources and Evaluation*, pages 1–26.

A Transnational Analysis of News and Tweets about *Nuclear Phase-Out* in the Aftermath of the Fukushima Incident

Philipp Heinrich[1], **Christoph Adrian**[2], **Olena Kalashnikova**[3], **Fabian Schäfer**[3], **Stefan Evert**[1]
Friedrich-Alexander-University Erlangen-Nuremberg
[1]Chair of Computational Corpus Linguistics, Bismarckstr. 6, 91054 Erlangen
[2]Chair of Communication Science, Findelgasse 7, 90402 Nuremberg
[3]Chair of Japanese Studies, Artilleriestr. 70, 91052 Erlangen
{philipp.heinrich, christoph.adrian, olena.kalashnikova, fabian.schaefer, stefan.evert}@fau.de

Abstract

Taking the impact of the Fukushima incident on the global discourse about nuclear energy as a case study, the present paper shows how to integrate computational linguistic methods into corpus-based discourse analysis (CDA). After an extensive literature review with regards to the related hermeneutic work, we present the corpus linguistic methods and point out methodological extensions. These extensions include visualization techniques that might help hermeneutic researchers explore large corpora, and second-order collocates, which help triangulate the semantics of lexical items. In our case study, we firstly give an in-depth analysis of the discourses that have formed around salient lexical items, in particular *nuclear phase-out* and *energy transition*, in the German *Frankfurter Allgemeine Zeitung* (FAZ) and the Japanese *Yomiuri* (both are conservative newspapers of the respective countries). We then provide preliminary results for the impact that the discourse had on German Twitter data. Last but not least, we show what effect the discourse has had on second-order collocates of the lexical item *Germany* in the FAZ corpus.

Keywords: Corpus-Based Discourse Analysis, Fukushima, Computer-Mediated Communication, Visualization, Collocations

1. Introduction

The world is currently witnessing a fundamental transformation of the political public sphere on a global scale. On the one hand, one can observe a powerful process of politicization in the guise of increasingly successful populist and oftentimes anti-European or even anti-democratic parties. On the other, some of the largest global protest movements against nuclear power, free-trade agreements, or social and economic inequality have occurred since the 1970s. In this regard, *Fukushima* represents a prototypical example for the intramedial and transmedial connections between national media discourses, namely in (1) the public spheres as represented in the mass media ("edited mass communication") and (2) the semi-public sphere in social media ("mass self-communication").

We consider the emergence of a global media discourse on renewable and nuclear energy in the aftermath of the Fukushima Daiichi disaster to be an ideal and politically highly relevant case study to study these multiple facets of this transformation of the transnational public sphere, particularly the shift in attitudes and opinions towards nuclear energy. This is not only due to the fact that it was a truly transnational global media event covered by the mass media worldwide (Hartwig et al., 2016; Hayashi, 2013; Ito, 2012; Koch et al., 2016), but also because it was an impetus for the formation of a transnationally networked anti-nuclear social movement (Kobayashi, 2011; Liscutin, 2011; Obinger, 2015; Sano et al., 2012; Slater et al., 2012; Thomson et al., 2012; Tsuda, 2011).

In our research, we analyze the tempo-spatial dissemination and framing of media discourses across different media and languages (i.e. their culture-specific linguistic realizations) in a complex connective network, which enables us to detect distinct distributional and linguistic patterns of the multiple forms of manipulation in the transnational algorithmic public sphere. In the paper at hand we present preliminary results of our approach to study the multidirectionality and intermedia complexity of the transformation of the transnational public sphere through a transmedial and transnational analysis of the articulated (verbal) connectivities in the media discourse on nuclear energy in Japan and Germany in the mass media and a social media network (Twitter[1]) in the years 2011–2014.

In particular, we will take multiple levels into consideration: (a) the intra-media process of dissemination of statements, attitudes, and opinions framing nuclear energy, (b) the inter-media reciprocities and convergences between social media and the mass media (e.g. trending topics on Twitter becoming news, test-marketing of headlines online before ending up in the printed version of a paper, or the dissemination of journalistic news via Twitter), (c) both of these on a temporal (i.e. 2011–2014) and a transnational axis (i.e. Germany and Japan).[2]

The two countries have been chosen because they have comparable mass media systems (public and private broadcasting, newspapers with a broad spectrum of conservative and liberal papers). Moreover, the two countries were affected by the Fukushima disaster in significantly different ways, which makes them ideal for contrastively studying transnational processes. It is usually argued that the accident turned into a global watershed for domestic atomic en-

[1]Twitter has turned into one of the most important tools in election campaigns and political agitation, aside from Facebook (Conway et al., 2015; Davis et al., 2016; Jungherr et al., 2016). Twitter is considered a particularly rich source of social media data because it has several advantages over other social media networks: accessibility, metadata availability, sample size, and the briefness of tweets (Burghardt, 2015; Mejova et al., 2015).

[2]The present work is based on one newspaper from each country as well as German Twitter data. Future research will include more newspapers and Japanese Twitter data.

ergy policies and public attitudes towards the use of nuclear power and the use of renewable energy. However, while in Germany the so-called "Fukushima Effect" has contributed to an immediate phase-out plan, in Japan itself, nuclear energy still remains a core element in the domestic energy mix. Moreover, with Japanese and German, our study covers western and non-western languages, offering not only promising results from linguistic and intercultural comparisons but also poses certain difficulties when studying distant languages in comparison or relation.

2. Related hermeneutic work

The global impact of the second largest nuclear accident in human history is commonly described as the Fukushima Effect in recent publications focusing on the short-term and long-term economic, political, or sociocultural consequences of the disaster (Arlt and Wolling, 2016; Hartwig et al., 2016; Hindmarsh and Priestley, 2016; Wolling and Arlt, 2014; Zeh and Odén, 2014). However, existing studies looking into the so-called Fukushima Effect in the sense of a transnational media discourse (e.g. the coverage and framing of the event itself or detectable changes in opinions and attitudes towards nuclear energy; cf. Scheufele (1999)) remain one-dimensional in terms of the methods or theoretical approaches applied, the types of media taken into consideration, and with regard to geographical regions under study, with a particular bias on Germany and Japan. They analyze attitudes and public opinion either across different media but only for a single national context, or from a transnational perspective but then only with regard to a single medium (usually newspapers).

Moreover, data sets cover a relatively short time period (usually from 4 weeks to at most 12 months after March 2011). Wolling and Arlt (2014) for instance, drawing on an analysis of German mass media coverage and survey data, diagnose a moderate Fukushima effect on Germany's political decision to phase out nuclear energy in summer 2011, arriving at the conclusion that the effect of Fukushima was much more drastic than that of Chernobyl because Japan was considered a high-tech country with tight security standards. Nienierza (2014), as well as Seiffert and Fähnrich (2014) arrive at a similar conclusion based on a content analysis of newspaper articles, and emphasize that nuclear energy had already been framed rather negatively in Germany prior to 2011.

Kepplinger and Lemke (2015) carried out an international and multimedia study of TV news and newspapers (but not social media) in the immediate weeks after the tsunami and meltdown, arguing that news coverage of Japan was related to domestic nuclear energy policy significantly more often in Germany and Switzerland (which opted for a phase-out as well), than in France or the UK, where nuclear energy still remains a core element of the domestic energy mix. Similarly, Hayashi (2013) studied German TV news reporting and found that 40% of the reports about the events in Fukushima were connected to Germany's domestic nuclear policy (cf. Ito (2012) and Honma (2016) for a critical perspective). Abe (2015), based on a content analysis of newspaper editorials from Mar 2011–Dec 2012, divides the Japanese newspaper landscape into a denuclearization camp (Asahi, Mainichi) and a pro-nuclear policy camp (Sankei, Yomiuri); Yoshino (2013) conducted a comparatively smaller study for Asahi and Yomiuri, but arrives at similar conclusions. Based on content and sentiment analysis of newspaper articles from Mar–Sept 2011, they found that while left-leaning Asahi frequently referred positively to Germany's phase-out, the conservative paper Sankei criticized Germany's dependency on its neighbors to sustain its energy demands in the transition period after the phase-out and emphasized Japan's economic vulnerability.

This bias in the reporting of many newspapers in Japan is also supported by a study by Satoh (2011) (see also Schäfer (2012); Schäfer (2017)), who analyzed the very limited reporting of anti-nuclear demonstrations in Tokyo (Mar–Sept 2011). Hartwig et al. (2016, 114), diagnosing a "gap in studies analyzing international news in Japan concerning Germany's energy policy shift after Fukushima", studied the Japanese mass media coverage (newspapers) of nuclear energy in Germany to understand the potential impact of what Gono'i (2015) has called a "boomerang [Fukushima] effect" on domestic nuclear energy policies in Japan. One might thus argue that the Fukushima Effect in the mass media and on social media worked in at least two directions, namely as a (re-)framing of nuclear energy inside and outside of Japan after the accident on the one hand, and as the reverse effect of this (re-)framing outside of Japan on the Japanese media discourse on the other.

3. Methodology

We touch related methodological work in the present section. However, it should be noted that we are striving for a methodological revision of discourse analysis, which is why there is only a very limited selection of related work at hand.

3.1. Keywords, Collocations, and Discourse

One of our goals is to give an in-depth analysis of the language data produced by mass and social media after the Fukushima incident. We biuld on corpus-based discourse analysis (CDA) (see e. g. Baker (2006); Baker et al. (2008)), which in principal boils down to the aggregation and subsequent deconstruction of concordance lines. The categories in which the textual data is divided in a CDA have to be made up by the hermeneutic researcher while dealing with the data, and cannot be known a priori. This approach thus differs fundamentally from automatic topic modelling or classification into ad hoc categories.

We understand discourses to be formed around lexical items, which we call key words, key items, or simply (discourse) nodes. The collocations of these key items are interpreted as attitudes or stances that can be taken towards the nodes. An example for a discourse is given by Baker (2006, 86): "refugees as victims."[3]

[3]Note that key words in this context does not necessarily mean words that have been determined by a keyword analysis in a traditional (corpus) linguistic sense. In fact, keyword analyses and collocation analyses can be translated into one another: a classical keyword analysis, which compares frequency lists of two corpora against one another, is equal to a collocation analysis where the context equals the respective texts of the keyword analysis.

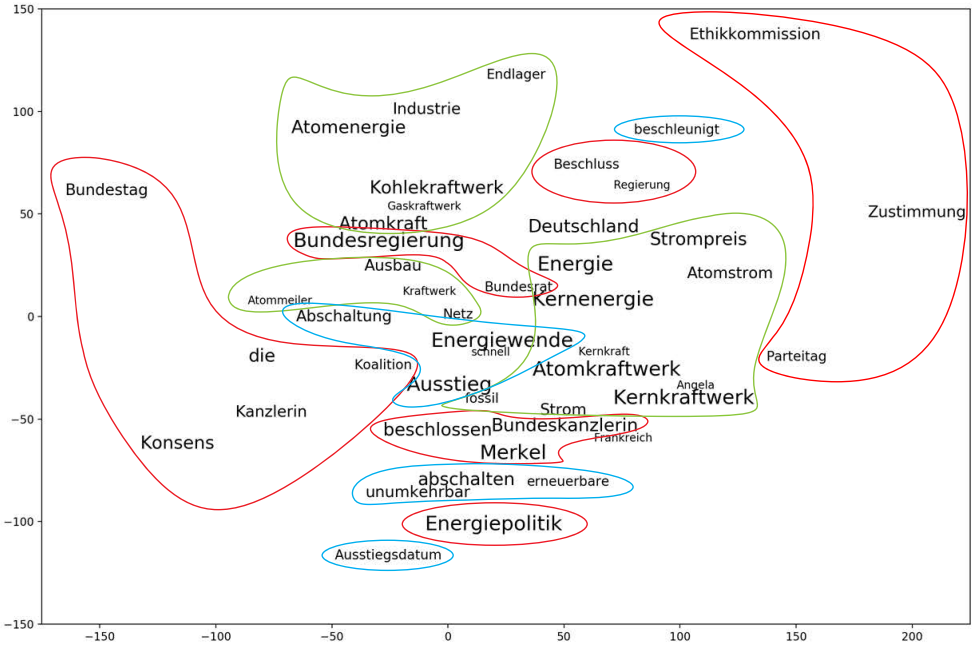


Figure 1: Paragraph-based collocates of *nuclear phase-out* (Atomausstieg) in the FAZ corpus of 2011, automatically arranged by semantic similarity using word embeddings and t-SNE. Size represents association strength to the node word (*nuclear phase-out*). Semantic fields are already loosely clustered together, which makes arrangement easier for hermeunitic interpretation. The post-hoc categories identified by the hermeneutic interpreter are *politics* (red), *actions and challenges of energy transition* (blue), and *economic capital and issues* (green).

In a collocation analysis, the window size (together with the association measure) influences the collocates that can be retrieved. A very conservative window size of, say, 3 will retrieve lexical items closely attributed to the node word, whereas wider window sizes might retrieve connotated items. Besides window-based collocations, we use generalized contexts such as tweets, paragraphs, and whole articles in the present paper. In our analyses, we use the standard log-likelihood measure as implemented in CQPweb (see Hardie (2012) for implementation details).

We do not use automatic Word Sense Disambiguation (WSD) techniques since we are interested in a more fine-grained triangulation of the semantic space: methodologically commited to distributional semantic models, we assume that the semantics of a word depends on the specific contexts at hand. It is thus reasonable and straightforward to look at second-order collocates when diving deeper into the semantics of a word, i. e. to retrieve items that collocate with the lexical node in the context of a discourse (see section 5.). Automatic WSD techniques, in contrast, can only categorize homographs into pre-defined (and very coarse) categories.

Although CDA reduces research bias – since the researcher is confronted with empirical data – this triangulation comes with an intense work load. We therefore actively develop (and use) visualization techniques for semantic fields, which make the results of keyword and collocation analyses (which are essentially n-best lists which have to be sorted by semantic categories) easier to interpret. On that account, we project word vectors (Mikolov et al., 2013) produced by parsing Wikipedia data onto a two-dimensional space using t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008). For an outcome of this process, see figure 1, which displays collocates of the lexical node *nuclear phase-out* in a newspaper corpus. Semantically similar words are automatically grouped together; the hermeneutic interpreter thus can rely on pre-grouped words to form categories.[4]

3.2. Corpora and corpus preparation

We use several newspaper corpora and data from social media, particularly Twitter, in our research. In the present paper, we focus on the *Frankfurter Allgemeine Zeitung* (FAZ) for Germany, and the *Yomiuri* for Japan (we have acquired corpora of both newspapers for the years 2011–2014), both being very conservative newspapers in the respective country. Additionally, we use a sample of 10,266,835 German

[4]Note that we can only display words for which vectors have been created. To this end, words that do not appear on Wikipedia frequently enough are not displayed at all.

tweets collected in the years 2011–2014. We also have a sample of roughly 300,000,000 Japanese tweets which we will analyze in further reserach.

The newspaper corpora are pre-processed with standard tools (TreeTagger[5] for German and MeCab[6] for Japanese). The TreeTagger software lemmatizes German tokens; MeCab splits Japanese texts into short-unit morphemes.

Processing computer-mediated communication (CMC) is more difficult: we use specialized software for tokenization (Proisl and Uhrig, 2016) and POS-tagging (Proisl, 2018) of German tweets. Lemmatization of CMC data is work in progress; to this end, we use TreeTagger to lemmatize German CMC. Moreover, building on earlier work, we deduplicate the data, which helps getting rid of social bots and other unwanted noise (Schäfer et al., 2017).

3.3. Measuring Impact

In the paper at hand, we measure the impact of an event (the Fukushima incident) and therefore know a priori what discourses we are looking for (and what kind of reaction we can expect): the incident had an obvious adverse effect on the public stance about nuclear energy, and anti-nuclear movements have thus formed in social media and have been seized by the mass media on the one hand; on the other, social media was influenced by news reports in the mass media; and discourses have subsequently found their way into politics. We hope that once we have analyzed these particular discourses, we can proceed to use the characteristics to give more information to the hermeneutic interpreter.

Note that key words and their collocates are suitable linguistic items for measuring impact: the statistical significance of collocates indicates the salience of the discourse at hand; the propagation of the found discourses (consisting of node words and their respective collocates) through the social network gives insights into the reception among the agents in the network.

4. Case study *Fukushima Effect*

We focus on the key word *nuclear phase-out* (脱原発) for the Japanese Yomiuri in the years 2011 and 2014. We first extract article-based collocations (see section 3.1.) of this keyword in section 4.1. We then show in section 4.2. how the discourse in German mass media was influenced by the Fukushima incident. In particular, we look at paragraph-based collocates of *nuclear phase-out* (Atomausstieg) and the more prevalent discourse around the topic node *energy transition* (Energiewende) in the German corpus of the FAZ.

Section 4.3. concludes the case study with a temporal analysis of the propagation of the lexical items through the German Twitter network. In section 5., we then show how the *nuclear phase-out* debate has spilled over to the distributional semantics of the lexical item *Germany* in the FAZ corpus.

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[6] https://taku910.github.io/mecab/

4.1. Analysis of Japanese mass media

In general, the change in discourses in the Yomiuri can be analyzed by performing a keyword analysis of the subcorpus of all the articles of 2014 compared to the subcorpus of 2011 (or vice versa). Such an analysis retrieves obvious items such as the lexical items *Tōhoku earthquake* and *tsunami*, *Fukushima accident*, *evacuation*, *radiation*, and *Khadafi* for 2011, compared to *Abe*, *Olympics* (*Sochi*, *Brazil*), *Ukraine*, *law-evading drug*, *Nankai megathrust earthquakes*, *consumption tax*, *reinterpretation of Japan's constitution*, *collective self-defense right*, *Mount Ontake eruption*, and *Islam* for 2014. We will leave it to the reader to come up with interpretations for this general keyword analysis.

An article-based collocation analysis of *nuclear phase-out* retrieves insights into the discourse surrounding the topic that we are interested in. In 2011, the top 50 collocates can roughly be categorized in three major semantic fields. Firstly, political actors and politics itself: *Kan* (菅), *Noda* (野田), *prime minister* (首相), the *democratic party* (民主党), *administration* (政権), as well as *policy* (政策), *transition* (転換), and *strategy* (方針). Secondly, the collocates are economic issues connect with the nuclear energy sector: *nuclear energy* (原子力), *power generation* (発電), *electricity* (電力), *operation* (稼動), and *reconstruction* (復興). Thirdly, collocates in the technological category that touch negative aspects of the nuclear energy sector: *safety* (安全), *fuel* (燃料), *shutdown* (停止), and *location* (立地).

In 2014, *nuclear phase-out* collocates on an article-level mostly with elections and politics (*point of dispute* (争点), *claim* (主張), *appeal* (訴える,訴え), *zero* (ゼロ), *policy* (政策), *street speech* (演説, as used in 街頭演説)) and the various political actors *Koizumi* (小泉), *Masuzoe* (舛添), *Hosokawa* (細川), *Abe* (安倍), the *Japan Restoration Party* (維新), and *councillor* (議員). There are fewer words regarding economics in comparison to 2011: *operation* (稼動), *energy* (エネルギー), *power generation* (発電), *electricity* (電力), and the economic policies advocated by Shinzō Abe (*Abenomics*, アベノミクス).

The differences in the *nuclear phase-out* discourse can thus be summarized in the following way: Firstly, the political actors changed. Secondly, not only was the Japanese administration discussed in 2011, but also political decisions of European leaders (Berlusconi, Sarkozy, Merkel, and Schröder). In 2014, political actors and organisations advocating phase-out and criticizing Abenomics appear, and discussions transfer to the street, which is suggested by the collocate *(street) speech* ((街頭) 演説); *nuclear phase-out* is thus used in political campagains. Among the speakers cited in the Yomiuri are former prime ministers Koizumi and Hosokawa, who became active supporters of nuclear phase-out. Debates in the economic sector were concerned on the restart of nuclear power plants after the inspections and critics towards Abenomics.

4.2. Analysis of German mass media

Table 1 gives an overview of the FAZ in the years 2011 and 2014 as well as the the discourses about *nuclear phase-out* and *energy transition* on a token- and paragraph-level. The salience of the *energy transition* debate has arguably not

restriction	#occurrences	per mill.	#articles	#occ. in par.	#par.	#tokens in par.
nuclear phase-out	1,051	7.05	748	919	844	104,121
– in 2011	525	16.08	409	525	473	57,828
– in 2014	103	2.95	88	97	93	10,753
energy transition	7,166	48.07	3,808	6,354	5,422	621,178
– in 2011	1,474	38.11	840	1,272	1,133	130,319
– in 2014	1,308	41.90	805	1,308	1,129	131,880

Table 1: Sub-corpora restricted to lemma searches for two key words. Left hand side: descriptive figures for the whole corpus (the whole corpus consists of 306,580 articles divided into 1,598,208 paragraphs with 149,058,904 tokens running text). Right hand side: descriptive figures for paragraphs.

much changed between 2011 and 2014: the lexical item *energy transition* (Energiewende) appeared 7,166 in 3,808 different articles in the whole corpus (amounting to 48.07 occurrences per million words); in 2011, it appeared 38.11 times per million words, compared to 41.90 times per million in 2014.[7]

The right hand side of table 1 shows the distribution of the respective items in paragraphs; *nuclear phase-out* e. g. has appeared 919 times in 844 different paragraphs[8], and these paragraphs amount to 104,121 tokens. When looking at the lexical item *nuclear phase-out*, we are diving deeper into a discourse about *energy transition*, since the former is one of the collocates of the latter. Note that the discourse about *nuclear phase-out* had become very prominent in the aftermath of the Fukushima incident in 2011, and had subsided by 2014.

The top 50 paragraph-based collocates of *nuclear phase-out* both in 2011 and in 2014 can roughly be divided into three categories just like in Japan (cf. figure 1 keeping in mind footnote 4): Firstly, political actors (Merkel, *federal chancellor* (Bundeskanzlerin), *federal government* (Bundesregierung)) and political issues such as *ethics commission* (Ethikkommission), *lifetime extension* (Laufzeitverlängerung), and *electricity supply* (Stromversorgung). Secondly, economic actors (German energy producers such as RWE, Eon, and EnBW) and the *price of electricity* (Strompreis). Thirdly, technological issues, such as *nuclear*, *gas-fired*, and *coal power plant* (Gaskraftwerk, Kohlekraftwerk, Atomkraftwerk) as well as *atomic mile* (Atommeiler) and the *electricity grid* (Stromnetz).

The technolocial aspects, political issues and actors have remained the same both in 2011 and 2014, presumably because the issues of a *nuclear phase-out* are associated with specific aspects, actions (for example the shutdown of nuclear power plants), the concept of *energy transition* and the challenges it presents. However, a deeper analysis of the collocates and concordance lines reveals some differences. In 2011, the FAZ mainly covered political actors, the energy industry and issues related to energy policy (*security of supply* (Versorgungsicherheit), *nuclear fuel taxation* (Brennelementesteuer)). Political actors debated about issues relating to the previous decision to prolonging the lifetime of nuclear power plants and the imminent decision to phase out nuclear energy. In 2014, the discourse shifted slightly towards collocates such as *investor protection* (Investorenschutz) and potential *lawsuits* (Klage).[9]

4.3. Analysis of German social media

Figure 2 shows the impact that the Fukushima incident had on discussions in the German Twitter network. Two preliminary notes: firstly, the top panel shows the monthly signal strength, which varies substantially over time. Whereas we have rather good coverage in 2011, we only have roughly one tenth of the data available for subsequent years.

Secondly, whereas the dashed lines show the salience of the topics in terms of a relative proportion of the number of tweets containing the topic node compared to all available tweets, the solid line shows the results when only taking into account originals (i. e. no retweets), which have furthermore been cleansed from any duplicates[10]. The effect of retweets and duplicates (some of this effect presumably being produced by social bots) can be interpreted in terms of "echo chambers", amplifying the original signal. Note that the figure displays a decrease in relative frequency when moving from the dashed to the solid line and is not due to an decrease of data material.[11]

We will restrict ourselves to two conclusions in the scope of this paper: Firstly, the *nuclear phase-out* debate sparked immediately after the incident and subsided rapidly afterwards. Secondly, although the *energy transition* debate also received a stark impetus on March 11, 2011, the debate seems to be independently prominent: it is a sustained topic that is being sparked many times in the aftermath. These increases in relative frequency can be used to extract points in time that indicate external incidents or events.

[7] This means that the discourse about this topic was more prominent in the years in between.

[8] The remaining occurrences can be attributed to titles, subtitles, and figure captions.

[9] The story appears similar when looking at article-based collocates; however, as expected, the collocates are more general when retrieved on the basis of articles. They then e. g. comprise the *TTIP*, and the midterm elections in the US. The *nuclear phase-out* debate thus seems to have been covered in the context of other debates.

[10] The deduplication takes place on heavily normalized tweets (Schäfer, 2017).

[11] In the case of the deduplicated data, the number of tweets containing the topic node as well as the denominator ("available tweets") is decreased.

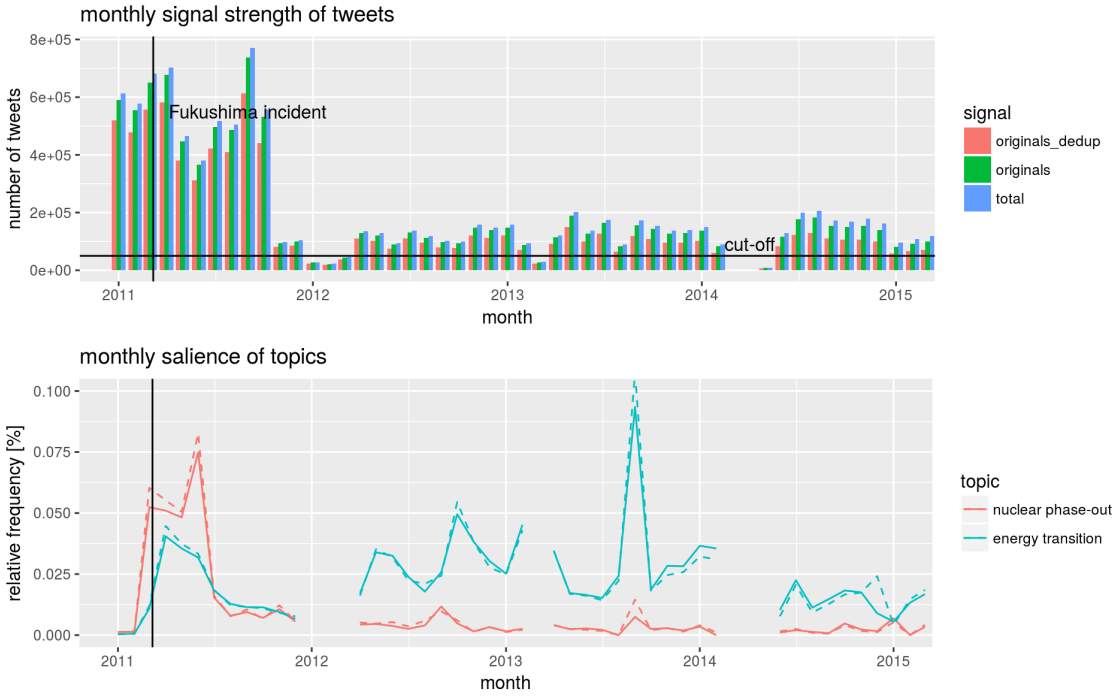


Figure 2: Top panel: Signal strength of our tweet data set (on a monthly basis). For our analyses, we only take into account months for which we have at least 50,000 tweets (as indicated by a horizontal line). Bottom panel: Temporal propagation of key words through the German Twitter network. The vertical line represents the time of the Fukushima incident.

5. Second-order collocates

We conclude by looking at second-order collocates[12] of the very salient[13] lexical item *Germany* in the FAZ corpus. Figure 3 shows collocates of *Germany* in the whole corpus (left hand panel) and in the sub-corpus of all paragraphs containing the lexical item *energy transition*. Recall that the visual arrangements is due to breaking down word vectors into two dimensions; the salience of the collocate determines its size. Note that the arrangement is the same in both settings (we use the same projection in both cases), but the size and selection of collocates varies.

In the unrestricted case, the most prominent collocates of *Germany* are realted to the Europe and its countries (*federal republik* (Bundesrepublik), *Europe* (Europa), *Austria* (Österreich), *Great Britain* (Großbritannien), and *France* (Frankreich)), as well as *Kabel* and *Alternative* (two very frequent multi-word units are "Kabel Deutschland", the largest cable television provider in Germany, and "Alternative für Deutschland", a right-wing political party that has gained popularity in recent years and has even made it into the German federal parliament in 2017) and *church* (Kirche), *evangelic* (evangelisch), *Jew* (Jude), *Muslim* (Muslim), and *Turk* (Türke). These collocates are relatively stable over time and can both be found in 2011 and 2014.

In the *energy transition* paragraphs, however, other collocates become salient, such as *future* (Zukunft) and *human being* (Mensch), alongside the more obvious items *energy transition* (Energiewende) *nature conservation* (Naturschutz), *(industrial or business) location* (Standort). What can be seen here is thus a spill-over effect of the energy transition debate into the very meaning of *Germany*, putting more weight on environmental issues and eliminating others such as religious and ethnic communities.

6. Conclusion and future research

Taking the *Fukushima Effect* as a case study, the present paper has shown how to marry computational linguistic resources with corpus linguistic methods. It points out several improvements over existing corpus linguistic techniques, in particular visualization by means of breaking down word vectors into two dimensions, and the definition of second-order (discourse) collocates. From a hermeneutic stance, it has shed light on the discourse that has been sparked after the Fukushima incident and that has propagated through the social network.

We report work in progress. In a next step, we will analyze further discourse nodes and connect the results from

[12]For ease of implementation, we use window-based collocations in the present study. The window-size is set to 5, and log-likelihood has been chosen as association measure.

[13]Albeit we already have a very large corpus of newspapers, looking at second-order collocates is only possible if the number of joint occurrences of discourse node and discourse collocate is high. Infrequent items such as *nuclear phase-out* are thus not yet suitable to be inspected.

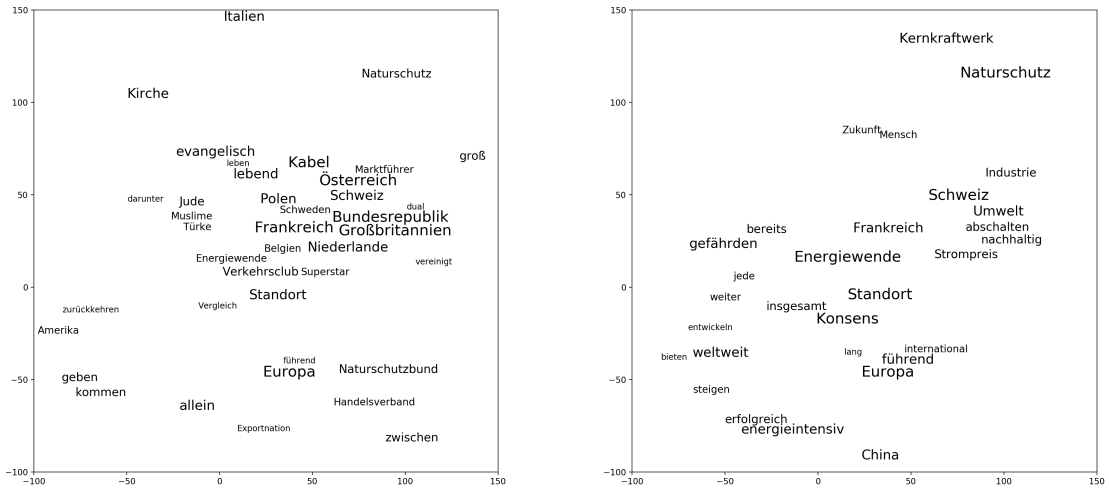


Figure 3: Collocates of the lexical item *Germany* (Deutschland) in the FAZ. Left hand panel: collocates when using the whole corpus as data basis. Right hand panel: second-order-collocates of *Germany* in energy-transition paragraphs.

across German and Japanese newspapers as well as the Twitter network. Further research is additionally necessary to make the methodology consistent (especially with regards to window-, paragraph-, or article-based collocates). Moreover, the Japanese Twitter data will be made usable by means of MeCab with a specialized CMC dictionary (Sato et al., 2017), and specialized word vectors will be created from Japanese Wikipedia and German and Japanese CMC data.

In the long run, we will also extend our data basis to more newspapers (particularly a left-wing newspaper from each country), and we will take into account behavioral operator-based connectivities for Twitter data (namely non-verbal actions like *retweeting* and *replying* to tweets).

7. Bibliographical References

Abe, Y. (2015). The nuclear power debate after Fukushima: a text-mining analysis of Japanese newspapers. *Contemporary Japan*, 27(2), January.

Arlt, D. and Wolling, J. (2016). Fukushima effects in Germany? Changes in media coverage and public opinion on nuclear power. *Public Understanding of Science*, 25(7):842–857, October.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3):273–306.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum, London.

Burghardt, M. (2015). Introduction to Tools and Methods for the Analysis of Twitter Data. *10plus1: Living Linguistics*, 1.

Conway, B. A., Kenski, K., and Wang, D. (2015). The Rise of Twitter in the Political Campaign: Searching for Intermedia Agenda-Setting Effects in the Presidential Primary. *Journal of Computer-Mediated Communication*, 20(4):363–380, July.

R. Davis, et al., editors. (2016). *Twitter and elections around the world: Campaigning in 140 characters or less*. Routledge, New York.

Gono'i, I. (2015). 2015-nen ANPO, Minshushugi wo futatabi hajimeru wakamono-tachi (ANPO in 2015. The Youth that is restarting Democracy), September.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Hartwig, M., Okura, S., Tkach-Kawasaki, L., and Kobashi, Y. (2016). Identifying the "Fukushima Effect": Assessing Japanese Mass Media Coverage of International Nuclear Power Decisions. *Journal of International and Advanced Japanese Studies*, 8:109–124.

Hayashi, K. (2013). Kiwadatsu doitsu no genpatsujikohōdō: Fukushima genpatsu jiko no kokusai hikakukenkyū yori: Tokushū 3.11 Fukushima dai'ichi genshiryoku hatsudensho jiko no meguru shakaijōohaō no kenshō: terebi jānarizumu, sōsharu media no tokusei to kadai (Highlighting Germany's nuclear accident news reporting: From an international comparative study on news reporting about the Fukushima accident (Special issue on the verification of social information environment covering the 3.11 Fukushima Dai'ichi Nuclear Power Plant Accident: Characteristics and challenges of TV journalism and social media). *Gakujutsu dōkō*, 18(1):50–55.

R. A. Hindmarsh et al., editors. (2016). *The Fukushima effect: a new geopolitical terrain*. Number 29 in Routledge studies in science, technology and society. Routledge, New York, NY.

Honma, R. (2016). *Genpatsu puropaganda (Nuclear Propaganda)*. Iwanami, Tōkyō.

Ito, H. (2012). Fukushima dai ichi genpatsu jiko ikō no genshiryoku hōdō – jiko-go 3-kagetsu-kan no shinbun shasetsu no ronchō kara miete kuru koto (The Atomic Energy News after the Accident at Fukushima No. 1 Nuclear Power Plant Viewed it from the Tone of Newspapers Editorials during the period Three Months after the Accident). *Pūru gakuin daigaku kenkyū kiyō*, 52:199–212, December.

Jungherr, A., Schoen, H., and Jürgens, P. (2016). The Mediation of Politics through Twitter: An Analysis of Messages posted during the Campaign for the German Federal Election 2013: The Mediation of Politics Through Twitter. *Journal of Computer-Mediated Communication*, 21(1):50–68, January.

Kepplinger, H. M. and Lemke, R. (2015). Instrumentalizing Fukushima: Comparing Media Coverage of Fukushima in Germany, France, the United Kingdom, and Switzerland. *Political Communication*, pages 1–23, June.

Kobayashi, A. (2011). *Saigai to social media: konran, soshite saisei e to michibiku hitobito no 'tsunagari' (The Catastrophe and Social Media: Confusion and the Connectivity that Led to Regeneration)*. Maikomi shinsho, Tōkyō.

Koch, M., Meyer, H., Nishiyama, T., and Zöllner, R. (2016). *Media-Contents und Katastrophen: Beiträge zur medialen Verarbeitung der Großen Ostjapanischen Erdbebenkatastrophe*. Iudicium Verlag.

Liscutin, N. (2011). Indignez-Vous! 'Fukushima,' New Media and Anti-Nuclear Activism in Japan. *Asia-Pacific Journal*, 9(47 (1)), November.

Yelena Mejova, et al., editors. (2015). *Twitter: A Digital Socioscope*. Cambridge University Press, New York.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nienierza, A. (2014). Die größte anzunehmende Umbewertung? Eine Frame-Analyse der deutschen Presseberichterstattung über Kernenergie nach den Reaktorunfällen von Tschernobyl (1986) und Fukushima (2011). In Jens Wolling et al., editors, *Fukushima und die Folgen – Medienberichterstattung, Öffentliche Meinung, Politische Konsequenzen*, pages 31–54. Univ.-Verl., Illmenau.

Obinger, J. (2015). *Alternative Lebensstile und Aktivismus in Japan: der Aufstand der Amateure in Tokyo*. Springer, Wiesbaden.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, et al., editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Proisl, T. (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*.

Sano, M., Varga, I., Kazama, J., and Torisawa, K. (2012). Requests in Tweets During a Crisis: A Systemic Functional Analysis of Tweets on the Great East Japan Earthquake and the Fukushima Daiichi Nuclear Disaster. In *39th ISFC*, pages 135–140, Sydney, Australia, July.

Sato, T., Hashimoto, T., and Okumura, M. (2017). Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing.

Satoh, K. (2011). What Should the Public Know?: Japanese Media Coverage on the Antinuclear Movement in Tokyo between March 11 and November 30, 2011.

Scheufele, D. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122, March.

Schäfer, F., Evert, S., and Heinrich, P. (2017). Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism and PM Abe Shinzō's Hidden Nationalist Agenda. *Big Data*, 5:1 – 16.

Schäfer, F. (2012). Fukushima: Rumours, Emotions and Rousseau's General Will in the Digital Age.

Schäfer, F. (2017). *Medium als Vermittlung: Medien und Medientheorie in Japan. (Medium as Mediation: The Media and Media Theory in Japan*. Springer, Wiesbaden.

Seiffert, J. and Fähnrich, B. (2014). Vertrauensverlust in die Kernenergie. Eine historische Frameanalyse. In Wolling, Jens et al., editors, *Fukushima und die Folgen – Medienberichterstattung, Öffentliche Meinung, Politische Konsequenzen*, pages 55–74. Univ.-Verl., Illmenau.

Slater, D. H., Nishimura, K., and Kindstrand, L. (2012). Social Media, Information and Political Activism in Japan's 3.11 Crisis. *Asia-Pacific Journal*, 10(24 (1)), June.

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., Isochi, R., and Wang, Z. (2012). Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter. In *9th Int. ISCRAM Conf.*, pages 1–10.

Tsuda, D. (2011). Sosharu media wa Tōhoku o saisei kanō ka? Lokaru komyuniti no jiritsu to fukkō (Can Social Media Revitalize the Tōhoku Region? Autonomy and Reconstruction of Local Communities). *Shisō chizu beta*, 2:52–73.

van der Maaten, L. and Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Jens Wolling et al., editors. (2014). *Fukushima und die Folgen - Medienberichterstattung, Öffentliche Meinung, Politische Konsequenzen*. Universitätsverlag TU Ilmenau, Ilmenau.

Yoshino, Y. (2013). Datsugenpatsu, hangenpatsuk ōdō ni kansuru shinbunkiji no sōi: Asahi shinbun to yomiuri shinbun (The Differences between Newspaper Articles on the Anti-Nuclear Power Movement. Asahi Shinbun and Yomiuri Shinbun). *Chikushi Jogakuen University*, 8:89–100.

Zeh, R. and Odén, T. (2014). Energieträger in der Berichterstattung. Die Nachwehen von Fukushima in

Schweden und Deutschland. In Jens Wolling et al., editors, *Fukushima und die Folgen – Medienberichterstattung, Öffentliche Meinung, Politische Konsequenzen*, pages 211–232. Universitätsverlag TU Ilmenau, Ilmenau.

Text and Graph Based Approach for Analyzing Patterns of Research Collaboration: An analysis of the TrueImpactDataset

Drahomira Herrmannova[†], Petr Knoth[‡], Christopher Stahl[†], Robert Patton[†], Jack Wells[†]
[†]Oak Ridge National Laboratory; [‡]The Open University
[†]Oak Ridge, TN, USA; [‡]Milton Keynes, UK
[†]{herrmannovad; stahlcg; pattonrm; wellsjc}@ornl.gov; [‡]petr.knoth@open.ac.uk

Abstract

Patterns of scientific collaboration and their effect on scientific production have been the subject of many studies. In this paper, we analyze the nature of ties between co-authors and study collaboration patterns in science from the perspective of semantic similarity of authors who wrote a paper together and the strength of ties between these authors (i.e. how frequently have they previously collaborated together). These two views of scientific collaboration are used to analyze publications in the TrueImpactDataset (Herrmannova et al., 2017) (Herrmannova et al., 2017), a new dataset containing two types of publications – publications regarded as seminal and publications regarded as literature reviews by field experts. We show there are distinct differences between seminal publications and literature reviews in terms of author similarity and the strength of ties between their authors. In particular, we find that seminal publications tend to be written by authors who have previously worked on dissimilar problems (i.e. authors from different fields or even disciplines), and by authors who are not frequent collaborators. On the other hand, literature reviews in our dataset tend to be the result of an established collaboration within a discipline. This demonstrates that our method provides meaningful information about potential future impacts of a publication which does not require citation information.

Keywords: collaboration networks, publication impact, text mining, semantic similarity, semantometrics

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan[1].

1. Introduction

Many studies have focused on scientific collaboration networks (Newman, 2004), patterns of scientific collaboration across disciplines (Friedkin, 1980), and on how these patterns affect scientific production and impact (Guimerà et al., 2005). Within this area, it has been shown that newcomers in a group of collaborators can increase the impact of the group (Guimerà et al., 2005), and that high impact scientific production occurs when scientists create connections across otherwise disconnected communities from different knowledge domains (Lambiotte and Panzarasa, 2009).

Existing works studying scientific collaboration networks have often focused either on properties of the network or on topical information pertaining to the nodes in the network. In this work we develop an approach which combines both network and topical information about the nodes. In order to gain insight into the types of collaboration between authors, we investigate the possibility of utilizing semantic distance in co-authorship networks together with the concept of *research endogamy* (Montolio et al., 2013) – the tendency to collaborate with the same authors or within a group of authors; and study how these types of collaboration reflect scientific importance.

In contrast to previous studies combining topical and network information (Glenisson et al., 2005; Janssens et al., 2006), our approach is beneficial in that it does not require citation information or a complete network, and can therefore be applied to newly published works. This approach, which we have introduced in a previous publication (Herrmannova et al., 2017), belongs to a class of methods referred to as "semantometrics" (Knoth and Herrmannova, 2014). In contrast to the existing metrics such as bibliometrics, altmetrics or webometrics, which are based on measuring the number of interactions in the scholarly network, semantometrics build on the premise that full-text is needed to understand scholarly publication networks and the value of publications. In this work we test our approach on a dataset of publications regarded as seminal and publications regarded as literature reviews by field experts, and compare these two publication types in terms of collaboration patterns.

2. Related Work

In this section, we review previous literature relevant to our study. First, we discuss methods for measuring the strength of ties in academic social networks, particularly research endogamy. Next, we briefly discuss methods for detecting communities in scholarly networks.

2.1. Strength of Ties in Academic Social Networks

Uncovering and studying patterns of academic social networks has been applied to many problems ranging from identifying influential researchers (Fu et al., 2014) and ranking conferences (Silva et al., 2014) to measuring re-

[1] http://energy.gov/downloads/doe-public-access-plan

search contribution (Rocha and Moro, 2016) and the diffusion of innovation (Valente, 1996). One of the first studies focusing on the strength of ties in social networks (Granovetter, 1973) introduced the concept "weak ties", i.e. ties across rather than within different communities or groups, and discussed the importance of these ties for diffusion processes. The tactic used to measure the strength of the tie between two individuals has in this case been to measure the proportion of common ties shared by the two individuals (Granovetter, 1973). Other approaches used to measure the strength of ties have been the frequency of contact (Granovetter, 1983), mutual acknowledgement of contact (Friedkin, 1980), or the likelihood of a tie re-appearing in the future (Brandão et al., 2017). (Newman, 2004) has proposed a measure of closeness of two authors which combines information about how many papers two authors wrote together and the number of other collaborators with whom they wrote them.

Following the ideas of (Granovetter, 1973) and later (Guimerà et al., 2005), who classified agents in a network as incumbents and newcomers, and have shown newcomers to a group help to improve its performance, (Montolio et al., 2013) have used the degree of new collaborations to rank conferences. The degree of new collaborations has been quantified using a new indicator called "research endogamy", which captures the inclination of a group to usually collaborate together. (Montolio et al., 2013) have shown the reputability of computer science conferences is correlated with the endogamy of their authors – low endogamy (i.e. less frequent collaboration) tends to be associated with highly reputed conferences, while lower quality conferences tend to publish articles by authors who have collaborated together on many occasions. (Silva et al., 2014) have applied the concept of endogamy to ranking publications and patents, and have shown low endogamy publications tend to receive more citations.

Overall, the aforementioned studies demonstrate the importance of connections across communities, diverse collaborations, and newcomers to a group. These patterns tend to be associated with high impact academic production. Hence, in this work, we use the concept of research endogamy of publications as defined by (Silva et al., 2014) to measure the strength of collaboration of a group of authors.

2.2. Semantic Similarity for Community Detection

Two approaches commonly used to detect communities in academic social networks are: (1) using the graph structure of the network or (2) using textual information of the nodes, e.g. by calculating semantic similarity between the nodes (Ding, 2011). These two approaches have also been used together to create maps of scientific communities in a specific field (Glenisson et al., 2005; Janssens et al., 2006) and to identify similar researchers (Cabanac, 2011). However, the network-based approach poses a significant challenge. Community detection in incomplete networks is a challenging task which requires the use of non-traditional methods (Lin et al., 2012). However, the complete network may not always be available, or may be difficult to obtain. For example, in order to identify whether two authors are members of the same community or of different communities, complete information about each of their communities (other authors and links between them) are needed.

Furthermore, network-based community detection has been shown to result in communities which span diverse topics, while text-based community detection helps in detecting nodes focusing on a specific topic (Ding, 2011). As we are interested in studying individual publications for which we may not have complete neighborhood information, we chose the text-based approach, and use semantic distance (the inverse of similarity) to measure the similarity of authors. This is also beneficial, as the textual similarity provides information complementary to the endogamy measure, which is calculated using topological information. By combining these two approaches, we are able to study collaboration networks not only from the perspective of tie strength, but also from the perspective of whether each tie represents potential knowledge transfer within or across disciplines.

3. Approach and Dataset

In (Herrmannova and Knoth, 2015), we have proposed a classification of research publications in which publications are divided into four groups (Figure 1) according to the semantic distance and the strength of ties between the publications' authors. In this paper, we provide an evaluation of this approach. To do this, we use the recently released TrueImpactDataset (Herrmannova et al., 2017) (Herrmannova et al., 2017) which contains publications of two types, seminal publications and literature reviews, and compare the collaboration patters of these two types of publications in terms of author distance and collaboration frequency.

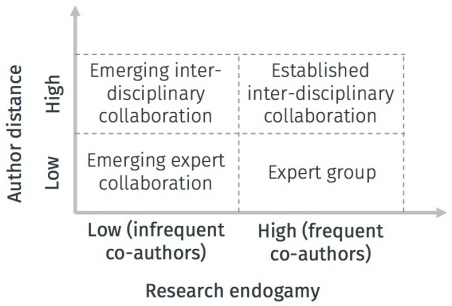


Figure 1: Types of research collaboration based on semantic distance of authors, and their collaboration frequency.

The semantic distance of a pair of authors is calculated using their previous publication record.

$$d(p) = \frac{1}{|A(p)| \cdot (|A(p)| - 1)} \sum_{a_i \in A(p), a_j \in A(p), a_i \neq a_j} d(a_i, a_j) \quad (1)$$

Here $A(p)$ is a set of authors of publication p. As explained in (Herrmannova and Knoth, 2015), we calculate the distance for a pair of authors $d(a_i, a_j)$ by concatenating the publications of each author into a single document. While this is a very simplistic approach, it is also beneficial in terms of complexity of the calculation.

In order to measure the strength of ties between authors, we combine the semantic distance with research endogamy value of the publication. Research endogamy (Montolio et al., 2013) is the tendency to collaborate with the same authors or within a group of authors. The research endogamy of a publication is calculated based on research endogamy of a set of authors A, which is defined similarly as the Jaccard similarity coefficient (Montolio et al., 2013; Silva et al., 2014) (Equation 2). The research endogamy $e(A)$ of a set of authors is calculated as follows:

$$e(A) = \frac{|\bigcap_{a \in A} P(a)|}{|\bigcup_{a \in A} P(a)|} \quad (2)$$

Here $P(a)$ represents a set of papers written by author a. Higher endogamy value is related to more frequent collaboration between authors in A – a value of 1 means all authors in A have written all of their publications together. On the other hand, a group of authors who have never collaborated together will have an endogamy value of 0.
Endogamy of a publication p is then defined as a mean of endogamy values of the power set of its authors (Montolio et al., 2013; Silva et al., 2014) (Equation 3).

$$e(p) = \frac{\sum_{x \in L(p)} endo(x)}{|L(p)|} \quad (3)$$

Here $L(p)$ is the set of all subsets with at least two authors of p, $L(p) = \bigcup_{k=2}^{k=|A(p)|} L_k(p)$, where $L_k(p) = C(A(p), k)$ is the set of all subsets of $A(p)$ of length k.

3.1. Methodology

To study the relation between author distance and research endogamy we use our TrueImpactDataset (Herrmannova et al., 2017), a multidisciplinary dataset of research publications containing seminal publications and literature reviews. We are interested in how these two types of papers are situated with regard to author distance and research endogamy. We use the following methodology. For the publications in the dataset we collect and/or calculate the following measures: (1) author distance, (2) research endogamy, (3) collaboration category (assigned to publications using author distance and research endogamy, Figure 1), (4) total number of citations per publication, (5) number of citations normalized by number of authors, and (6) number of citations normalized by publication age. To compare seminal publications and literature reviews in our dataset with regards to author distance and research endogamy we use t and χ^2 tests to determine whether the values of the measures are statistically significant for seminal publications and literature reviews. To analyze whether author distance and research endogamy help in distinguishing between seminal publications and literature reviews in our dataset we also analyze the distributions of both features and the placement of seminal publications and literature reviews within the four collaboration categories (Figure 1).

3.2. Data

To collect all data needed for studying the measures introduced in Section 3., we have used three data sources:

1. TrueImpactDataset[2] (Herrmannova et al., 2017) (Herrmannova et al., 2017), which provides us with seminal publications and literature reviews,
2. Microsoft Academic (MA) API[3] (Sinha et al., 2015) which we use to collect metadata (particularly the information about authors and their publications) of the papers in the TrueImpactDataset,
3. Mendeley API[4] which we use to collect publication abstracts.

Table 1 shows the size of the dataset. After collecting all needed data the size of the original dataset was reduced to 144 publications (i.e. publications for which we were able to obtain author information) – 75 literature reviews and 69 seminal publications. The row *Number of authors* shows the total number of (non-disambiguated) authors of all papers in the dataset.

Publications in TrueImpactDataset	314
TrueImpactDataset publications in MA	298
Pubs with author information in MA	144
Number of authors	758
Total number of publications	27,653

Table 1: Dataset size. The table shows for how many of the TrueImpactDataset publications we managed to get the needed metadata and how many additional publications we collected (i.e. including all other publications of the authors in the TrueImpactDataset – row *Total number of publications*).

4. Experiments

In this section, we investigate how seminal publications and literature reviews are situated with regard to the extracted features. To do this, we use the following methodology: we take all of the 144 core papers and for each of them collect the features defined in section 3.1.. To understand whether seminal publications and literature reviews differ in terms of these features we calculate an independent one-tailed t-test for each feature except for the collaboration category feature which is categorical and for which we calculate χ^2 test. The t-test is a measure commonly used to assess whether two sets of data are statistically different from each other. In other words, it helps to determine the features that can distinguish survey papers from seminal papers. To test the significance, we set the significance threshold at 0.05. Furthermore, for each feature we create a histogram and by comparing these histograms for the two publication types we gain insight into norms and placement of seminal and survey publications in terms of metrics.
The complete results of the t-test are presented in Table 2 and the histograms for the five numerical features are shown in Figure 2. For four of the features we reject the

[2] `trueimpactdataset.semantometrics.org/`
[3] `aka.ms/academicgraph/`
[4] `dev.mendeley.com/`

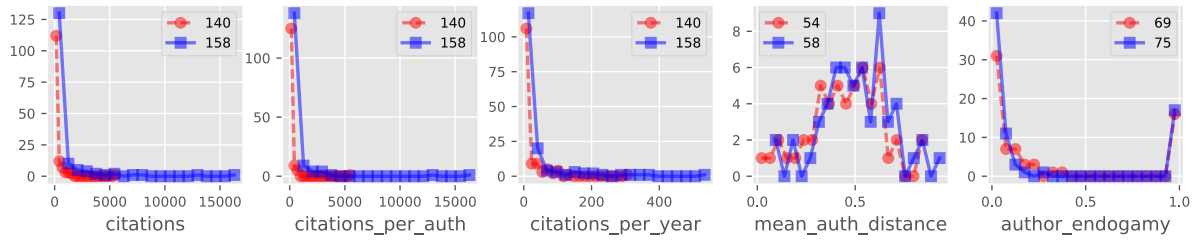


Figure 2: Histograms of the five numerical features.

null hypothesis of equal means. The t-test tells us the values of these four features are significantly different for the two sets of papers.

Metric	p**-value**
Mean author distance	0.0327
Endogamy	0.3217
Citations	0.0012
Citations per year	0.0073
Citations per author	0.0110
Collaboration category	0.0218

Table 2: Results of t- and χ^2 tests.

Next, we analyze the collaboration category feature which is assigned to publications using the values of author distance and research endogamy (Figure 1). We calculate χ^2 test, which is a statistical test for categorical variables for testing whether the means of two groups are the same, to test whether the seminal publications and literature reviews differ in terms of the collaboration category. The resulting p-value is 0.0218 (Table 2), which is lower than our significance threshold of 0.05. This tells us that the means of the two sets of papers differ.

Figure 2 shows the endogamy values for the dataset are strongly skewed towards 0. Furthermore, the results of the t-test suggest research endogamy by itself does not distinguish between the two publication types. However, when combined with the author distance measure, a clear pattern emerges, which is visible in Figure 3. Figure 3 shows the relation between author distance and research endogamy, represented as the number of publications belonging to each collaboration category introduced in Figure 1. To create this figure, we have first assigned each publication two values – its author distance and research endogamy. We have then used median endogamy (0.0297) and median author distance (0.4996) to separate the publications in the dataset into the four categories presented in Figure 1.

The figure shows there are some differences between seminal publications and literature reviews. In particular, the main difference between the two classes is that emerging collaborations (i.e. when the authors have not collaborated frequently together previously) are in our dataset more common for seminal publications. On the other hand, literature reviews seem to be a result of established collaborations within a discipline. These observations are consistent with previous studies which have shown that cross-

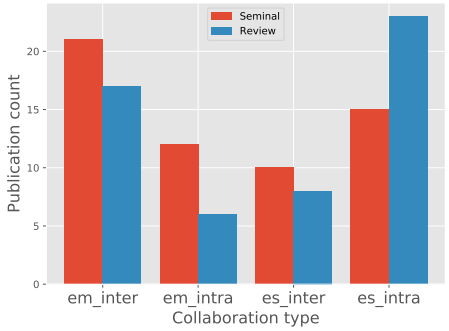


Figure 3: Number of publications belonging to each collaboration category across both publication types.

community citation and collaboration patterns are characteristic for high impact scientific production (Guimerà et al., 2005; Lambiotte and Panzarasa, 2009; Montolio et al., 2013). We believe this is an encouraging result which suggest semantic distance of authors combined with their endogamy value might be helpful in providing early indication of future impacts of a publication.

5. Conclusions

This paper studied the relationship between semantic distance of authors which collaborated on a publication and the strength of ties between these authors, which was assess using research endogamy measure (a measure of collaboration frequency introduced by (Montolio et al., 2013)). More specifically, we compared publications of two types – seminal publications and literature reviews – in terms of their author distance and research endogamy values. Our results show that there are distinct differences between these two publication types in terms of collaboration patters. In particular, we found that seminal publications tend to be written by authors who have previously worked on dissimilar problems (i.e. authors from different fields or even disciplines), and by authors who are not frequent collaborators (i.e. emerging inter-disciplinary collaborations). On the other hand, literature reviews in our dataset tend to be the result of an established collaboration within a discipline (an "expert group"). This demonstrates content analysis might provide valuable information for research evaluation and meaningful information about potential future impacts of a publication which does not require citation information.

6. Bibliographical References

Brandão, M. A., Vaz de Melo, P., and Moro, M. M. (2017). Tie strength persistence and transformation. *AMW (to appear)*.

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3):597–620.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498–514.

Friedkin, N. (1980). A test of structural features of granovetter's strength of weak ties theory. *Social networks*, 2(4):411–422.

Fu, T. Z., Song, Q., and Chiu, D. M. (2014). The academic social network. *Scientometrics*, 101(1):203–239.

Glenisson, P., Glänzel, W., Janssens, F., and De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41(6):1548–1572.

Granovetter, M. S. (1973). The strength of weak ties. In *American Journal of Sociology*, volume 78, pages 1360–1380. Elsevier.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233.

Guimerà, R., Uzzi, B., Spiro, J., and Nunes Amaral, L. A. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(April):697–702.

Herrmannova, D. and Knoth, P. (2015). Semantometrics in coauthorship networks: Fulltext-based approach for analysing patterns of research collaboration. *D-Lib Magazine*, 21(11/12).

Herrmannova, D., Patton, R. M., Knoth, P., and Stahl, C. G. (2017). Citations and readership are poor indicators of research excellence: Introducing trueimpactdataset, a new dataset for validating research evaluation metrics. In *Proceedings of the 1st Workshop on Scholarly Web Mining*.

Janssens, F., Leta, J., Glänzel, W., and De Moor, B. (2006). Towards mapping library and information science. *Information processing & management*, 42(6):1614–1642.

Knoth, P. and Herrmannova, D. (2014). Towards semantometrics: A new semantic similarity based measure for assessing a research publication's contribution. *D-Lib Magazine*, 20(11):8.

Lambiotte, R. and Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics*, 3(3):180–190.

Lin, W., Kong, X., Yu, P. S., Wu, Q., Jia, Y., and Li, C. (2012). Community detection in incomplete information networks. In *Proceedings of the 21st international conference on World Wide Web*, pages 341–350. ACM.

Montolio, S. L., Dominguez-Sal, D., and Larriba-Pey, J. L. (2013). Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, 42(2):11–16.

Newman, M. E. (2004). Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks*, pages 337–370. Springer.

Rocha, L. and Moro, M. M. (2016). Research contribution as a measure of influence. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2259–2260. ACM.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr., W., and Laender, A. H. F. (2014). Community-based Endogamy as an Influence Indicator. In *Digital Libraries 2014 Proceedings*, page 10.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P., and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social networks*, 18(1):69–89.

7. Language Resource References

Herrmannova et al. (2017). *TrueImpactDataset*. Distributed via `http://trueimpactdataset.semantometrics.org/`, ISLRN 197-407-228-291-9.

TSCC: a New Tool to Create Lexically Saturated Text Subcorpora

Zygmunt Vetulani, Marta Witkowska
Adam Mickiewicz University in Poznań
Poland
vetulani@amu.edu.pl, martusiazielinska@gmail.com
Umut Canbolat
University of Kocaeli
Turkey
u.canbolat@yahoo.com

Abstract

In the paper we present a new tool to evaluate lexical saturation of text corpora, where lexical saturation refers to a state in which it is hard to find new lexemes outside the corpus. Estimation of the saturation degree for a given corpus contributes in a natural way to the corpus quality evaluation. We propose saturation tests as a stopping criterion for subcorpora creation. Although the first application of the TSCC tool is the evaluation of lexical coverage of corpora, it may be equally useful to study corpora representativeness for various phenomena, and – more generally – their usefulness for corpus-based research, both theoretical and practical (as e.g. studies of information impact). It may serve for cost evaluation of expensive engineering tasks in language competence modelling for AI purposes as well as in literary research. The system (TSCC) is highly language independent, i.e. it may be applied directly or easily adapted to any language in which the text units may be represented in alphabetic scripts. Its preliminary version (OCASSC) has been tested on a corpus of clients' opinions published by booking.com. The prototype will be freely distributed for beta testing.

Keywords: text corpora, corpora quality, lexical saturation tests, subcorpora creation, stopping criterion

1. Introduction

Although there is consensus about fundamental importance of linguistic data corpora (texts, recordings) for investigating natural languages according to the world-observation-based methodology of natural sciences, there is still a need of commonly accepted methods for text corpora evaluation. Initially, the size of the corpus was considered as a sufficient quality measure for corpora but quickly it has become clear that this is not an absolutely effective solution for corpora quality evaluation. The need of producing linguistic models for particular applications brought the attention of language engineers to specific linguistic phenomena. Consequently, corpora for language modelling are supposed to be *representative* for the phenomena in question[1]. As corpus collection is expensive (time, effort) and difficult (legal issues), quality evaluation of existing corpora is an important issue.

Representativeness of corpora was largely discussed in the wider context of corpora quality sometimes opposed to the concept of size as quality measure. In the frequently cited paper Douglas Biber (1993) presented, without any formal definition however, what it means to 'represent' a language. He considered various aspects of this concept taking into consideration language data stratification, sampling and – last but not least – size. Biber's work provides argumentation in favor of both qualitative and quantitative basis for corpus design (Kennedy, 1998).

In this study we follow the observation that "a huge corpus is not necessarily a corpus from which generalization can be made. A huge corpus does not necessarily 'represent' a language or a variety of a language any better than a smaller corpus" (Kennedy, 1998). This observation seems to be particularly adequate in a study of local lexicographical phenomena for which the question "how many tokens of a lexical item are necessary for descriptive adequacy" (ibid.) is justified.

2. Corpus Saturation

There exist various methods to estimate the representativeness of a sample of data for a given phenomenon. What they have in common is the evaluation of the chance of finding a new manifestation of the phenomenon outside the sample. In a representative sample all relevant examples should occur at least once. If the sample is (almost) representative then (almost) all newly observed manifestations will be identical to some already done. In particular, the size of the list of single manifestations of the phenomenon (list of hapaxes) will decrease or remain the same after each new observation. The decrease in the number of manifestations results from the increase in the corpus saturation with respect to the considered phenomenon.

3. Lexical Saturation

We will explore the concept of lexical saturation of a corpus (cf. e.g. Kittredge 1983, also Vetulani 1989). This concept appeared useful in research on the evaluation of the size of virtual vocabulary of sublanguages[2] (e.g. in the context of machine translation) and was used to study lexical saturation of corpora. Informally, we say that corpus is lexically saturated when "new lexemes appear only sporadically as a result of the extension of the corpus in a natural way" (Vetulani 1989).

In order to study the lexical saturation we consider the corpus to be a linearly ordered set of elements (words, symbols etc.). For its initial segments of the length N we observe the number of different words V. This function is increasing and the observation of the growth of V informs us about the degree of saturation of the corpus. The

[1] We consider a corpus as representative for a given language phenomenon, or a class of phenomena if it contains examples for all relevant aspects of this phenomenon.

[2] See e.g. (Kittredge 1983).

function may be represented graphically by a saturation graph. Observations of corpora confirm that V grows systematically with N but slower. The reason for this is that the observed vocabulary becomes more and more saturated, i.e. it is more and more difficult to introduce new words into discourse (Vetulani 1989).

For a sound[3] data gathering procedure it is crucial to have a good stopping condition, i.e. criterion to stop data collection. A good stopping condition will prevent against collecting data beyond necessity.

Let us consider the corpus as a linearly ordered partition into segments of equal size. Observation of the number of new words ΔV in the last segment informs us about the degree of lexical saturation of the corpus. It follows that a sufficiently small value[4] of the ratio ΔV/ΔN is a good candidate for the stopping condition. Checking whether this stopping condition is satisfied is called *lexical saturation test*.

If we intend to compare saturation degree between corpora of different sizes it may be convenient to calculate the ratio ΔV/ΔN for the last segment representing X% of the whole corpus (denoted ΔV/ΔN(X%) and called (*X% growth ratio*) for all corpora (and then to compare).

This method may be generalized in a natural way to evaluate representativeness of a corpus with regard to various phenomena. For example, in order to evaluate the minimum size of a balanced opinion corpus we performed experiments involving opinion adjectives (as adjectives are – for most of languages – the main lexical tool to support classification of opinions into negatives or positives).

Notice. Lexical saturation as stopping condition may be inadequate for corpus based studies of some global phenomena where statistical methods demanding huge amount of data or neural algorithms requesting large training sets are in standard usage. (See e.g. McEnry and Hardie (2012) or Peris et al. (2017)).

4. OCASSC

In 2017 we implemented the system OCASSC (Opinion Corpora Acquisition Software for Subcorpora Creation) initially designed to create corpora of opinion texts. In particular, it was used to randomly generate possibly small subcorpora of a large collection of texts that could be considered representative for studies of lexical instruments to express opinions. For the purpose of investigation of corpora representativeness for the given phenomena, OCASSC system was equipped with a functionality that enables execution of incremental saturation tests.

OCASSC requires two sets of input data. The first one is a corpus in an XML format. Such a corpus may be generated by the system OCAS (Vetulani et al. 2015). The second set of data is a predefined list of elements whose role is to restrict the search space. These are words included in the so called *reference list* that may (but do not need) appear in the investigated corpus and are formal indicators of the relevant phenomena (for example, opinion adjectives, i.e. adjectives that may be used to support an opinion, as illustrated below). The program retrieves the input data to create subcorpora of the length specified by the user and to perform saturation test for the predefined reference list.

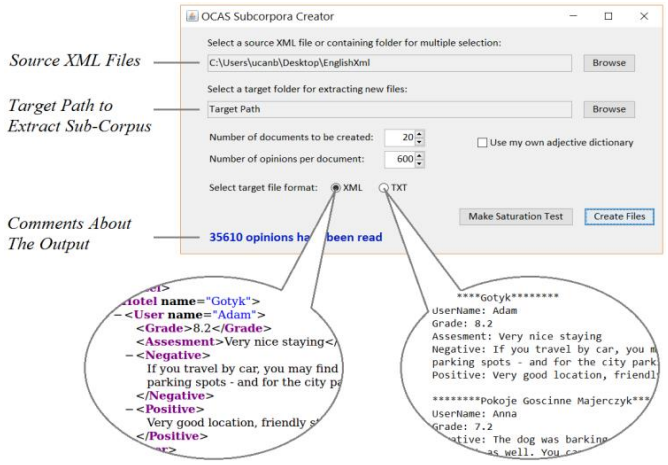


Figure 1: The OCASSC main screen

4.1 Input Data Examples

OCASSC accepts data in XML format where tags are used to separate the meta-information from the opinion texts. The (simplified) example provided below presents the input compatible with the one used by the system OCAS (Opinion Corpora Acquisition Software) for collecting opinions (Vetulani et al. 2015).

<All>
<Review>
<HotelName>Gotyk House</HotelName>
<Positive>The welcome, the location and the wonderful helpfullness and charm of the staff were notable. Breakfast was simple and ample. We chose to go in the cold (-10C) and were perfectly warm.</Positive>
<Negative>Would have liked a kettle in the room but accept that fire considerations in such an old house prevented it.Q</Negative>
</Review>
<Review>
<HotelName>Hotel Kazimierz</HotelName>
<Positive>The location was just what we wanted, the room was clean and quiet and the staff were friendly and helpful</Positive>
<Negative>nothing</Negative>
</Review>
<Review>
<HotelName>Hotel Polski Pod Białym Orłem</HotelName>
<Positive>Room was large for European standards and the bed was so comfortable. Breakfast was good with a broad array of

[3] Data gathering procedure is considered *sound* with respect to a given objective, if it guarantees acquisition of all data necessary to reach this given objective.

[4] The value is to be fixed depending on the purpose of the corpus design and development.

foods, they also had good scrambled eggs and sausages. Would definitely stay here again.</Positive>
<Negative>Nothing, it was perfect for our stay in this beautiful city.</Negative>
</Review>
</All>[5]
(We have underscored opinion adjectives)

4.2 Output Data Examples

OCASSC was designed to generate subcorpora with the desired properties (saturation), so the basic format of the resulting subcorpora is the same as input. However, for more readability the system may output data in a text format.

*******Gotyk House*******
Positive: The welcome, the location and the wonderful helpfullness and charm of the staff were notable. Breakfast was simple and ample. we chose to go in the cold (-10C) and were perfectly warm.
Negative: Would have liked a kettle in the room but accept that fire considerations in such an old house prevented it.
*******Hotel Kazimierz*******
Positive:The location was just what we wanted, the room was clean and quiet and the staff were friendly and helpful
Negative: nothing
*******Hotel Polski Pod Białym Orłem*******
Positive: Room was large for European standards and the bed was so comfortable. Breakfast was good with a broad array of foods, they also had good scrambled eggs and sausages. Would definitely stay here again.
Negatative: Nothing, it was perfect for our stay in this beautiful city.

4.3 Experiment

Configuration of the OCASSC system requires a priori definition of the search space for the phenomenon of concern. In our experiment the search space was determined by the list of all adjectives that may be used to express opinion and which we consider interesting for our purposes. To create this list we useed a corpus of 2040 opinions in English (for hotels in the city of Poznań). We proceeded to manual annotation of all occurrences of adjectives used as opinion words. The next step was to create a frequency list of all annotated adjectives. This list contained 490 various adjectives (for 11854 occurrences). 312 adjectives that occur more than once in the corpus were used in the experiment as reference lists (we discarded hapax legomena in order to limit the number of atypical opinion words in the reference list). Then we applied OCASSC to a corpus of opinions (in English) about hotels in Poland containing over 850.000 of text words (34.800 opinions containing 28.371 different words) in order to extract subcorpora of 2040 opinions. We applied the *10% growth ratio* (with respect to opinions) to evaluate the degree of saturation for these subcorpora. In each of the observed cases the ratio varied between 0.01 and 0.03.

On the other hand, application of OCASSC to the corpus of 34.800 opinions and to the set of *all words* as reference list, shows that the corpus is far from being lexically saturated as the *10% growth ratio* for the reference list containing all (general) vocabulary is relatively very high (0.453). Consequently, a significant growth of the observed vocabulary should be expected when proceeding to its extension.

5. TSCC – Text SubCorpora Creator

The purpose of the software is to create smaller corpora from a large text corpus. Reusing the OCASSC system architecture, we have designed and implemented (2017, summer) a system operating in open (unformatted) text as a tool for extracting subcorpora of any desired size from a large corpus. In situations where the representativeness of the corpus is closely related to its lexical completeness (which is the case of sublanguages determined by restricted application domains, such as e.g. the sublanguage of metereological reports[6]) evaluation of the degree of lexical saturation using TSCC may help to fix the stopping criterion for creation subcorpora with desired properties.

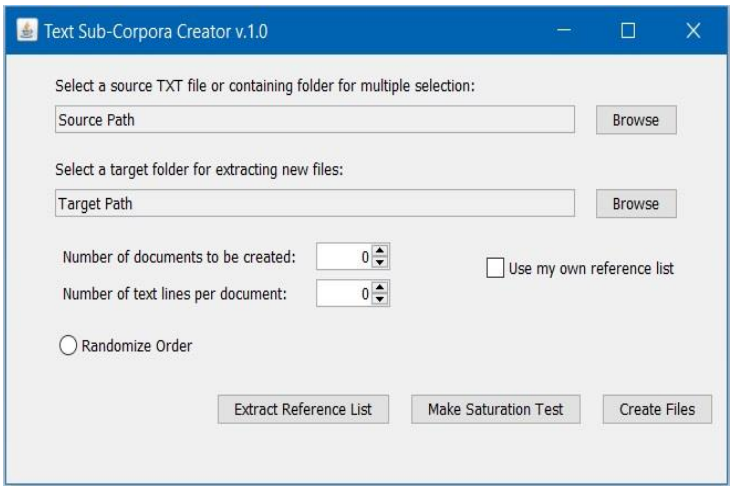


Figure 2: The TSCC main screen

5.1 TSCC v.1.0 Functionalities

The system processes the input corpus in the form of a text(.txt file). The "Source Path" browse button will permit the user to provide location and open the input text. The input text may have a form of just one or several input files contained in a folder. Selection of the source file or folder will result in evaluation of the whole text (corpus) and the total number of lines will be displayed. The next step is to select the target folder like we did for selecting the source

[5] Spelling and syntax are original.
[6] The studies of the sublanguage of metereological reports were the object of special interest in the classical linguistic and AI literature ((Muller (1975), Kittredge (1983)).

path. The output files will be put in a folder created by the system.

After the input/output operations, the remaining processing parameters must be declared. The corpus text will be considered as composed of text lines grouped into "documents" containing fixed number of lines. This number, as well as the number of documents are to be specified by the user. These two numbers will determine the length of the subcorpus extracted and processed by TSCC. Creation of documents may be done line by line (default solution) or in a random way with respect to the whole input corpus.

"Reference list" plays the same role as in OCASSC: it defines the set of elements (words or special tags) to be taken into account to define the saturation function (and therefore determine the processing search space). The user is supposed to use his/her own reference list ("Use my own reference list" checkbox).

Finally, the user is supposed to declare one of the following three functionalities: to create a subcorpus of the specified size (button "Create files"), to calculate data for the saturation test (button "Make Saturation Test"), or to transform the input corpus into a reference list composed of all word forms of the corpus[7] (button "Extract Reference list"). The results are being saved in the output folder.

5.2 Experiment

TSCC generates data necessary to draw the saturation graphs for text corpora (function "Make Saturation Test"). Figure 3 presents the graph for a corpus composed of 8920 lines randomly extracted from a larger corpus of hotel opinions (given by Booking.com clients). The corpus appears not large enough to be considered as lexically saturated (for adjectives) because the increase of numbers of adjectives is quasi linear after the first 4000 lines. Still it is not very high, as the local increase speed expressed by the 10% ratio equals 12 words per 1000 lines). (The 10% ratio with respect to the number of text words equals 0.0012).[8]

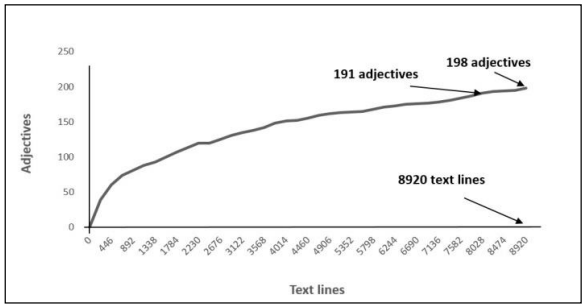


Figure 3: Saturation graph for all adjectives observed in the corpus of 8920 text lines (hotel opinions)

6. Possible Application in the Information Impact Studies

Measuring information impact is an important practical issue in many contexts: political, social (public security), military, etc. Impact is measurable as far as information is registered and stored in a systematic way, e.g. in form of text corpora. As it is a rule in empirical studies, appropriate sampling[9] is crucial. TSCC, or equivalent tools, may be useful in this venture but requires a careful selection of reference lists of words as information filters. Examples will be presented and discussed at the workshop. Below we limit ourselves to the main ideas only.

Impact of an event may be estimated on the ground of the registered information in the form of text. We will assume that information we are interested in is represented in a text corpus (called basic corpus), e.g. in a corpus of press news, and that the event description may be identified by a list (reference list) of terms (or concepts) (possibly including proper names, acronyms, dates etc.). The reference list must be individually defined for the events in question.

If an event may be identified (on the ground of the reference list) through a search in small, random selected samples (subcorpora) extracted from the basic corpus, (i.e. it is easy to find in the basic corpus) then we will be entitled to conclude that its impact is big. In practice, generation of the samples (subcorpora) may be done using the TSCC system for a predefined stopping criterion in order to guarantee the appropriate saturation of the generated samples by the elements of the reference list.

In a similar way we may evaluate what is the part of a well-defined subject area (with respect to other subject areas) in the literary output of an author. In that case the whole literary production of the author (or its representative fragment) constitutes the basic corpus, and lexical formal subject indicators form the reference list. It is in our imminent plans to apply this method to analyze the literary works of Polish writer and poet Julia Hartwig (https://en.wikipedia.org/wiki/Julia_Hartwig) and to present the results at the CIDTD Workshop at LREC 2018.

7. Further Research

We intend to further develop the TSCC system from the point of view of literary research. In particular its utility will be beta tested in the research on vocabulary structure of particular authors and particular literary works. We hope to prove its utility for stylometry. The beta prototype of the TSCC tool will be released for LREC 2018 and distributed under an open license.

[7]This may be useful in order to evaluate the degree of lexical saturation of the corpus with respect to the whole lexicon, e.g. using the X% ratio method described above.

[8]The Reader should however be warned that the saturation parameters in the experiments of sections 4.3 and 5.2 ought not to be confronted, as the reference lists (adjectives in both cases) are different. In particular, the used reference lists are different and the Princeton WordNet list of adjectives used in 5.2 is very poor in opinion-supporting adjectives.

[9]By *sampling* we mean here selection of subcorpora of appropriate size (small).

8. Acknowledgement

We intend to thank Jolanta Bachan, Katarzyna Klessa, Gerard Ligozat, Patrick Paroubek and Suleyman Menken for their precious assistance and help.

9. References

Biber, D. (1993). Representativeness in corpus design, *Literary and Linguistic Computing*, Vol. 8, Nb. 4, pp. 243–257.

Kennedy, G. (1998). *An Introduction to Corpus. Linguistics*. Longman: London and New York.

Kittredge, R. (1983). Semantic processing of texts in restricted sublanguage. *Computers & Mathematics with Applications*, Vol. 9, Issue 1, pp. 45–58.

McEnry, T., Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Muller, Ch. (1975). Peut-on estimer l'étendue d'un lexique? *Cahiers de Lexicologie*, Nb. 27, 1975–II, pp. 3–29.

Peris, Á., Chinea-Ríos, M., Casacuberta, F. (2017). Neural Networks Classifier for Data Selection in Statistical Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, Vol. 108, June 2017, pp. 283–294.

Vetulani, Z. (1989). *Linguistic problems in the theory of man-machine communication in natural language. A study of consultative question-answering dialogues. Empirical approach*. Brockmeyer: Bochum.

Vetulani, Z., Witkowska, M. and Menken, S. (2015). Corpus Based Studies on Language Expression of Opinions. In: Z. Vetulani and J. Mariani, editors, *Proceedings, 7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, 2015*. Fund. UAM: Poznań, pp. 365–369.

The Financial Attention Index to Measure Impact of Crisis from Microblog

Zhongsheng Wang, Kiyoaki Shirai
School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology

wzswan1.0@jaist.ac.jp, kshirai@jaist.ac.jp

Abstract

This paper proposes a new financial index called Financial Attention Index (FAI) to measure an extent of a financial risk. A stock market drastically changes when many novice investors participate in it. At the same time, they often posts messages on Social Networking Service. Therefore, the FAI is calculated as ratio of financial related comments on microblog to detect a financial crisis. Furthermore, we train a model to predict future stock prices from a history of stock values and the FAI. Results of experiments show effectiveness of our proposed index to capture an impact of a financial crisis.

Keywords: Financial Index, Financial Crisis, Social Networking Service, Machine Learning

1. Introduction

A financial crisis, such as a drastic drop of stock prices, may cause considerable losses to many investors. At the same time, due to globalization, a financial crisis in one country may influence stock markets across the whole world. Because a stock market is an innate complex, dynamic and chaotic, the management of financial risk has been proved to be a very difficult task. For many decades, researchers have tried to analyze historical stock prices or a company's financial statements to measure financial risk (Fama et al., 1969; Fama, 1991; Cootner, 1964). However, the results were still not quite helpful for risk management.

In the Socio-economic Theory of Finance (Prechter Jr et al., 2012; Prechter Jr and Parker, 2007), irrational speculation behaviors play an import role in a financial crisis. Social Networking Service (SNS), such as Twitter or Weibo, are now widely used by large numbers of people. These social networks provide us with a considerable amount of information that can be used to monitor the financial market and which can also be used to predict a financial crisis.

The goal of this present research is to propose a new financial index that measures the extent of a financial crisis. Microblogs are used as a source of our new index. The hypotheses behind our financial crisis index are as follows.

- Not many investors focus on financial markets daily. When a bull market begins and stock prices keep going up, more beginners come into the markets.
- These people are intense and they make irrational speculation decisions. This may cause a panic and this can lead to a bear market. The behaviors of beginners may make a financial risk more serious.
- When many novice investors come into the market, they post messages about financial topics on SNSs. Therefore, the intensity of attention towards finance on microblogs can have a positive correlation with the level of a financial bubble. In particular, the numbers of financial messages posted on SNSs can be used as a financial crisis index.
- This index can be used as an observable financial market singularity for risk management.

In this paper, we will also attempt to use our proposed financial crisis index to predict the financial market trend for risk management. Specifically, a model to predict future stock prices is trained with past stock prices and our index that is derived from texts posted on a microblog. Given that historical prices are sequential data, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have been used to predict a stock's price (Chen et al., 2015; Heaton et al., 2017). We train the LSTM and measure its predictive ability to evaluate the effectiveness of our proposed financial crisis index.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 explains our proposed financial index for risk management. Section 4 presents the training of the LSTM for prediction of stock values. Section 5 reports the results our experiments and evaluates our proposed method. Finally, Section 6 concludes this paper.

2. Related work

2.1. Financial analysis theory

Financial market analysis has been one of the most attractive areas of previous research. Many researchers have tried to interpret the current financial situation from several different academic perspectives. The econometrics theory model is the most famous and influential of these methods. In addition, many financial market studies have been based on Fama's Efficient Market hypothesis (Fama et al., 1969; Fama, 1991). Although this hypothesis considers that the current price of an asset always reflects all of the previous information available for it instantly, it is impossible for us to collect all of the necessary information to make a prediction of the future price.

The other famous economic theory is the Random-walk hypothesis (Cootner, 1964; Malkiel, 1973), which claimed that a stock price changed independently of its history. However, information other than historical prices can also be used for stock price prediction, such as the financial

news that is released every day. At the same time, this theory considers that it is impossible to predict a financial market. In contrast, many previous studies have already proven that the stock market only followed those theories during specific periods (Glantz and Kissell, 2013).

2.2. Use of textual data for financial market analysis

The previous studies have only worked on history data and past stock prices. However, many other factors can be taken into consideration when analyzing the market. For example, thanks to their widespread use, textual information from Web forums and SNSs can be used for market analysis.

Nguyen et al. proposed a method based on sentiment analysis on social media to predict the movement of stock prices (Nguyen and Shirai, 2015; Nguyen et al., 2015). A new topic model, called Topic Sentiment Latent Dirichlet Allocation (TSLDA), infers topics and their sentiments simultaneously and has been incorporated into the prediction model.

Jaramillo et al. proposed a method to predict a stock price using a history of prices, and also the polarity of company reports and news (Jaramillo et al., 2017). In this study, the polarity of the texts is identified by Support Vector Machine (SVM).

Jianhong et al. applied a deep learning method on sentiment-aware stock market prediction (Li et al., 2017). They tried to analyze sentiment of documents in a stock forum with a Naive Bayes model. They then trained an LSTM neural network to a predict stock value using the results of the sentiment analysis as an input.

Similar to these previous studies, we also use textual information for a stock market analysis. However, our main focus is not to predict a stock's value but instead to avoid financial risk.

2.3. Prediction of financial market and risk

Machine learning is often applied for financial market forecasting with historical data. These methods are important related work because this paper also applies machine learning for stock movement prediction with our proposed index.

For example, Nelson et al. used LSTM neural networks to predict stock market price movement (Nelson et al., 2017). In addition, several methods have proposed to apply deep learning for multivariate financial series (Batres-estrada, 2015; Heaton et al., 2017). Chen et al. used an LSTM-based method for China stock market return prediction (Chen et al., 2015).

Although most researchers and investors care about a good return in the financial market, managing risk can be more important because it can help to avoid massive loss. Some of the previous work on the financial risk management was based on historical prices only; however, several studies have also tried to use textual analysis.

For example, Niemira and Saaty proposed a method to build Analytic Network Process model for financial crisis forecasting (Niemira and Saaty, 2004). This paper trained a turning point model to forecast a financial crisis likelihood based on an Analytic Network Process framework.

Meanwhile, Oh et al. proposed a method to use neural networks to support early warning system for financial crisis forecasting (Oh et al., 2005). Using nonlinear programming, the procedure of DFCI (daily financial condition indicator) construction is calculated by integrating three sub-DFCIs, which are based on different financial variables.

Trusov et al. used company financial reports in a multi-representation approach to text regression of financial risks (Trusov et al., 2015). Finally, Kogan et al. also used financial report regression for financial risk prediction (Kogan et al., 2009). In particular, they used Support Vector Regression (SVR) as a prediction model.

Although these previous studies have tried to use textual information for financial risk management, texts on SNS were not given attention. In this paper, comments on Weibo are used as a source of textual information for risk management.

3. Financial Attention Index

3.1. Definition

We propose to use the Financial Attention Index (FAI) to measure the extent of a financial risk. The FAI is defined as in Equation (1).

$$FAI \stackrel{def}{=} \frac{\textit{number of financial related comments on SNS}}{\textit{total number of comments on SNS}} \quad (1)$$

As discussed in Section 1, we suppose that financial topics are to more mentioned and discussed on SNSs when many novice investors participate in the market, which may cause market instability. Therefore, the FAI is assumed to be positively correlated to a financial risk. Although a financial crisis can be measured from various points of view, the FAI can only be used a financial crisis index from one perspective.

Figure 1 shows the procedure to calculate FAI. The comments on SNS are classified to determine whether or not they are related to financial topics. In this study, we will use the Weibo. The number of financial comments and total comments are then counted to get the FAI. However, a classifier of financial related comments is not trained from Weibo comments and, instead, the labeled data of news articles is used for training.

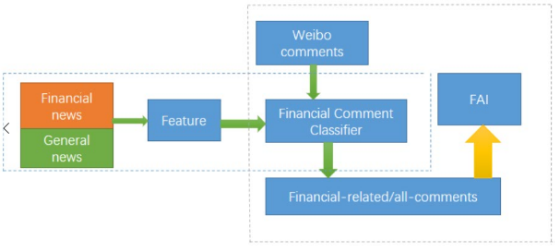


Figure 1: Procedure to calculate FAI

In our research, FAI is calculated for every week; the number of comments posted in a period of one week is counted to get FAI.

3.2. Calculating FAI

3.2.1. Type of classifier

Two kinds of financial classifiers are trained.

- Two-way classifier

  This classifies a comment on Weibo as a financial or non-financial related comment.

- Three-way classifier

  This classifies a comment on Weibo into three classes: (1) Stock-related, which is a comment related to the stock market, (2) Financial-related, which is a comment related to financial markets such as future market, bond market and so on, but not related to the stock market, and (3) Other, which is a comment not related to financial topics. Because we mainly focus on analysis of stock prices to detect a financial crisis, we distinguish topics about stocks with other financial related topics. When the three-way classifier is used for calculation of FAI, both stock-related and financial-related comments are treated as financial comments.

3.2.2. Data collection

Two kinds of data are collected to calculate FAI.

Weibo dataset

Comments on Weibo posted from 2013-7-1 to 2016-12-5 are crawled. This period contains stable, bull, and bear markets, as reported later. The number of collected comments is 2,104,746. They are collected with their posting time to calculate the FAI for each period of one week.

News dataset

News articles are collected from two sources: the Tencent news website and the text collection of THUC Project. Tencent includes the websites of general news[1] and financial news[2]. In the latter, news stories are categorized into several topics. The news in the topic category of "The New Third Board"[3] are used as stock-related documents, while the news stories in the other topic categories are financial-related documents. News stories in the general news website are crawled as other (non-financial) documents. The THUC text collection is developed by Tsinghua University (Sun et al., 2006). In this dataset, each document is annotated with its topic (stock related, financial related and other). Table 1 shows the number of documents in the news dataset.

The news dataset is used to train both the two-way and three-way financial classifiers. When the two-way classifier is trained, news articles of the stock and financial classes are treated as the financial related documents.

[1] http://news.qq.com/

[2] http://finance.qq.com/

[3] A name of stock market in China

Table 1: News dataset

Class	Website	THUC Project
Financial related	800,000	5,000,000
Stock related	5,000	200,000,000
Other	165,000	500,000,000

3.2.3. Training the classifier

SVM is used to train the financial classifier. SVM has been widely applied in the classification of documents, such as sentiment analysis. It is considered as the most appropriate learning algorithm for unbalanced datasets with a large number of features. A linear function is chosen as the kernel function of SVM. The gensim (Řehůřek and Sojka, 2010), sklearn (Pedregosa et al., 2011), scipy (Jones et al., 2001) and jieba (Rossum and Guido, 1995) [4] tools are used to train SVM.

Bag-of-words are used as features for training SVM. The bag-of-words model is a simplified representation of a document, which is widely used in natural language processing and information retrieval (McTear et al., 2016). In information retrieval, a weight of an index term is often determined by the TF-IDF (term frequency-inverse document frequency), which reflects how important the term is in a text collection (Jaramillo et al., 2017). In this study, the function words are removed by preprocessing. All content words in a document are extracted as the features. The weight of each feature is set as the TF-IDF score.

We found that the number of the features was high; that is, nearly 200,000. Therefore, we apply Latent Semantic Analysis (LSA) to reduce the feature space. LSA is used to reduce the size of a matrix of words by documents using Singular Value Decomposition (SVD). In this study, the number of the features is reduced to 50.

4. Stock price prediction with FAI

To evaluate its risk management ability, the FAI is used to predict a stock index. LSTM (Xiong et al., 2015) is chosen to train the prediction model because it is well used in the prediction of time series. In our model, the input of LSTM is a time sequence of either the stock index, a difference of the stock index, or FAI. The difference of the stock index is defined as a change of the stock index between the current and previous periods. We also train a model where these three kinds of values are concatenated as a vector and passed to the input layer of LSTM. The output of LSTM is a stock index of the next period. We define a period of LSTM as one transaction day. Note that FAI is calculated for each week. If the FAI is used for LSTM, then the same value is entered during days in a week. Our LSTM structure consists of one input layer, two LSTM layers and one output layer. The input and output layers consist on one neuron node. The first and second LSTM layers contain 5 and 100 nodes, respectively.

LSTM is learned through training by the Python deep learning library Keras (Chollet, 2015). The activate function in

[4] It is used for word segmentation of Chinese texts.

LSTM units is 'linear'. The model is trained by the rmsprop method with 1 example in a batch, with categorical cross entropy as the objective loss function. The validation fraction is set as 0.1%. The learning rate is set as 0.001. All of the initial weights are set to be small positive constant values. To prevent overfitting, a dropout is set at 20% and an L2 regularization constraint is set as 0.01.

5. Evaluation

5.1. Classification of financial comment

The classifier to judge whether a comment is related to financial topics takes an import role in FAI. First, the financial classifier is empirically evaluated by a 10-fold cross validation on our news dataset. The performance of the classifier is measured by the accuracy, which is defined as a ratio of the number of correctly classified comments to the total number of comments.

Table 2 shows the accuracy of the two-way and three-way classifiers. Recall that the comments are classified as either "financial-related" or "other", even when the three-way classifier is used, while both stock-related and financial-related comments are regarded as financial-related comments.

Table 2: Result of classification of financial comments

Model	Accuracy
Two-way classifier	83.35%
Three-way classifier	86.18%

The performance of the financial classifiers is satisfying. We found that the three-way classifier outperformed the two-way classifier and, therefore, the three-way classifier is used in our next experiments.

5.2. Correlation between FAI and the real stock index

To evaluate how well the FAI can work when used to predict a financial crisis, the correlation between FAI and a real stock index is measured. Given that our FAI is derived from comments of Weibo, which is a Chinese microblogging service, the SSE Composite Index (SCI) is chosen as the stock index in this experiment. The SCI is computed from the stock prices of Chinese companies. It is a tool that is widely used by investors to describe the market.

The SCI values are obtained from the Finance Sina website[5]. The dataset contains the SCI values of 679 trading days during 2013-10-28 to 2016-8-2, which is almost the same period where the comments on Weibo are downloaded.

The F-test, which is a common method for statistical test, is applied to measure the correlation between FAI values and the real SCI values for 679 trading days. We also apply the F-test between FAI and the difference of the stock index, in addition to the stock index and the difference.

Table 3 shows the results of the F-test. It can be seen that FAI strongly correlates with the stock index, but not with the difference of the stock index. It seems difficult to measure the change of the stock index using only FAI.

[5]http://finance.sina.com.cn/

Table 3: Result of F-test

Two variables	**F-test**
FAI vs. Stock Index	0.000111
FAI vs. Difference	0.0155
Stock Index vs. Difference	0.00718

5.3. Prediction of stock index movement

5.3.1. Experimental setting

Our LSTM-based stock index prediction model is evaluated for its ability to predict not the stock index but, instead, a movement of the stock index. More precisely, we consider a task to classify a movement of the stock index in one week into one of the following three classes.

Up : This is a case where the stock index goes up by T or more, as shown in (2),

$$P_e - P_s > T. \tag{2}$$

, where P_s and P_e are the stock index at the start and end of the period, respectively.

Keep : This is a case where the stock index does not change drastically as shown in (3),

$$|P_e - P_s| \leq T. \tag{3}$$

Down : This is a case where the stock index goes down by T or more, as shown in (4),

$$P_e - P_s < -T. \tag{4}$$

We set T as 0.02 in this experiment.

We have chosen 120 continuous weeks in our stock index dataset to be used as test periods. For each test period, all of the past values are used as the training data as shown in Figure 2. Note that more training data is available when the stock movement of the later week is predicted. For the first test period, data of previous 80 transaction days is prepared for training. We chose this experimental setting so that more data can be used for training the LSTM.

5.3.2. Evaluation criteria

Two evaluation criteria are used in this experiment:

Accuracy

This is a proportion of the number of the test periods for which the predicted movement class agrees with the true class.

No Lost Accuracy

The main goal of this research is not to help investors get a good return but to help them avoid financial risks. Consequently, it is more important to predict the movement "Down" to avoid making a loss. Therefore, we introduce another criterion, which is called "No Lost Accuracy". This is the accuracy of the stock movement prediction task where the test periods are classified as either "Down" or not. That is, "Up" and "Keep" classes are merged into one class: "Not-Down".

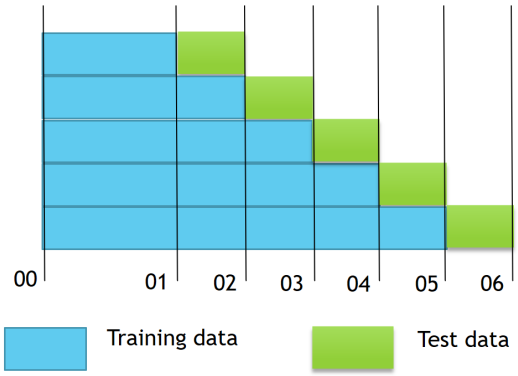


Figure 2: Test and training periods of prediction model

5.3.3. Result and discussion

Table 4 shows the Accuracy (A) and No Lost Accuracy (NLA) of four different prediction models. Here, "SI", "DI", "FAI" stand for the model using only the stock index, the difference of the stock index, and FAI, respectively. "All" stands for the model using three values.

Table 4: Results of the stock movement prediction

Model	SI	DI	FAI	All
A	31.1%	42.9%	41.2%	35.3%
NLA	47.9%	62.2%	54.6%	57.1%

From these results, it can be seen that "FAI" is better than "SI" in terms of Accuracy and No Lost Accuracy. This indicates that FAI is effective and is able to predict the movement of the stock index. When combining FAI with other indexes, the No Lost Accuracy is also improved ; however, "FAI" does not outperform "DI". Therefore, it is found that the difference of the stock index is a strong indicator of the movement of the stock market.

Figure 3 shows a change of SSE Composite Index in our dataset. Both a bull market and a bear market are found in this graph. To evaluate the performance of the prediction models in different situations (bull market, bear market etc.), we divide the test periods into 12 terms, where each term consists of 10 test periods (10 weeks), and we then measure an average of Accuracy and No Lost Accuracy on each term. The results are shown in Table 5. The tables include the situation, which is graphically shown in Figure 3, for each term.

The results of our experiment show a tendency for the "FAI" model to achieve better performance than the other models in both a bull and a bear market. In particular, it is good on a front bull situation (T5). In terms of a bear market (T8, T9, and T10), the "FAI" model is better than or comparable to the others. These results indicate that FAI is effective and it is able to predict a drastic movement of the market. In contrast, the model using the difference of the stock index (DI) works well for stable situations. The model using the stock index (SI) has the worst results in our experiment. Unexpectedly, the "All" model is not always the best choice. Although three values are simply combined as one input vector in our model, this might be a bit too naive.

Table 5: Prediction for different situations

(a) Accuracy

Term	Situation	SI	DI	FAI	All
T1	stable	50%	60%	40%	40%
T2	stable	10%	60%	50%	40%
T3	stable	30%	40%	40%	20%
T4	prim bull	30%	70%	10%	50%
T5	front bull	40%	30%	90%	50%
T6	mid bull	40%	30%	40%	40%
T7	late bull	20%	50%	40%	30%
T8	front bear	20%	30%	40%	0%
T9	mid bear	10%	40%	30%	50%
T10	late bear	40%	20%	50%	10%
T11	stable	30%	40%	40%	50%
T12	stable	44%	56%	22%	44%

(b) No Lost Accuracy

Term	Situation	SI	DI	FAI	All
T1	stable	70%	60%	50%	70%
T2	stable	10%	80%	60%	70%
T3	stable	70%	60%	50%	20%
T4	prim bull	50%	90%	40%	50%
T5	front bull	60%	50%	100%	80%
T6	mid bull	50%	50%	50%	70%
T7	late bull	30%	60%	40%	50%
T8	front bear	40%	50%	50%	50%
T9	mid bear	30%	60%	70%	60%
T10	late bear	50%	50%	50%	40%
T11	stable	50%	70%	50%	70%
T12	stable	67%	67%	44%	56%

6. Conclusion

This paper investigates whether the financial attention of investors as measured from a large collection of comments on Weibo could predict a down occasion on the SSE Composite Index (SCI). The FAI is defined as a proportion of the financial related comments and estimated by the financial classifier trained from the news articles. The FAI, in addition to the stock index and the difference of the index, were used as the input of the LSTM model to predict the future stock index.

The results of our experiment showed that the accuracy of the LSTM with FAI was no better than the model with the difference of the stock index on average but it was better or comparable in bear markets. Therefore, the FAI can be an effective index to predict a financial risk and this can help investors to avoid making a massive loss. In addition, the FAI can also effectively predict the beginning of a bull market because the prediction model with FAI worked well in a front bull term. That is, the FAI can be used to detect turning points in a stock market, no matter if prices move down or up.

Although the results of the experiments have proven the effectiveness of our proposed method, there is still room for

Figure 3: Change of SSE Composite Index

improvement. Given that the number of SNS comments might be insufficient, we plan to get more comments from Weibo to improve the quality of the FAI. Currently, three indexes (the stock index, the difference of the index, and FAI) are simply concatenated in our LSTM-based prediction model; however, a means to combine them should be explored in more detail. Consequently, in our future work, we will investigate the design of a structure of LSTM to accept multiple inputs.

7. Bibliographical References

Batres-estrada, G. (2015). Deep learning for multivariate financial time series.

Chen, K., Zhou, Y., and Dai, F., (2015). *A LSTM-based method for stock returns prediction: A case study of China stock market*, pages 2823–2824. Institute of Electrical and Electronics Engineers Inc.

Chollet, F. (2015). Keras. *Git-hub*.

Cootner, P. H. (1964). *The random character of stock market prices*. M.I.T. Press.

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, 10(1):1–21.

Fama, E. (1991). Efficient capital markets: II. *Journal of Finance*, 46(5):1575–617.

Glantz, M. and Kissell, R. (2013). *Multi-asset Risk Modeling, 1st ed.* Academic Press.

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 10.

Jaramillo, J., Velasquez, J. D., and Franco, C. J. (2017). Research in financial time series forecasting with SVM: Contributions from literature. *IEEE Latin America Transactions*, 15(1):145–153, Jan.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL ,09, pages 272–280.

Li, J., Bu, H., and Wu, J. (2017). Sentiment-aware stock market prediction: A deep learning method. In *2017 International Conference on Service Systems and Service Management*, pages 1–6, June.

Malkiel, B. G. (1973). *A Random Walk Down Wall Street*. Norton, New York.

McTear, M., Callejas, Z., and Griol, D., (2016). *The Dawn of the Conversational Interface*, pages 11–24. Cham.

Nelson, D. M. Q., Pereira, A. C. M., and Oliveira, R. A. d. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1419–1426.

Nguyen, T. H. and Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In *ACL*.

Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.*, 42(24):9603–9611, December.

Niemira, M. and Saaty, T. (2004). An analytic network process model for financial-crisis forecasting. *International Journal of Forecasting*, 20:573–587.

Oh, K. J., Kim, T. Y., Lee, H. Y., and Lee, H., (2005). *Using Neural Networks to Support Early Warning System for Financial Crisis Forecasting*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prechter Jr, R. R. and Parker, W. D. (2007). The financial/economic dichotomy in social behavioral dynamics: The socionomic perspective. *Journal of Behavioral Finance*, 8(2):84–108.

Prechter Jr, R. R., Goel, D., Parker, W. D., and Lampert, M. (2012). Social mood, stock market performance, and U.S. presidential elections: A socionomic perspective on voting results. *SAGE Open*, 2(4):2158244012459194.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings*

of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50.

Rossum and Guido. (1995). Python reference manual. *CWI (Centre for Mathematics and Computer Science)*.

Sun, M., Guo, Z., Zhao, Y., Zheng, Y., Si, X., and Liu, Z. (2006). Thu Chinese text classification.

Trusov, R., Natekin, A., Kalaidin, P., Ovcharenko, S., Knoll, A., and Fazylova, A. (2015). Multi-representation approach to text regression of financial risks. In *2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, pages 110–117.

Xiong, R., Nichols, E., and Shen, Y. (2015). Deep learning stock volatilities with google domestic trends.

Impact of Scientific Research beyond Academia: An Alternative Classification Schema

Andreas Witt°,■, Jana Diesner×, Diana Steffen°, Rezvaneh Rezapour×, Jutta Bopp°, Norman Fiedler°, Christoph Köller*, Manu Raster*, Jennifer Wockenfuß*

°Institut für Deutsche Sprache, Mannheim, Germany
{witt, bopp, fiedler, steffen}@ids-mannheim.de
■University of Cologne, Germany
andreas.witt@uni-koeln.de
×School of Information Sciences, University of Illinois at Urbana-Champaign, USA
{jdiesner, rezapou2}@illinois.edu
*The German National Library of Science and Technology, Hannover, Germany
Manu.Raster@tib.eu
*Görgen & Köller GmbH, Hürth, Germany
{j.wockenfuss, c.koeller}@gk-mb.com

Abstract

The actual or anticipated impact of research projects can be documented in scientific publications and project reports. While project reports are available at varying level of accessibility, they might be rarely used or shared outside of academia. Moreover, a connection between outcomes of actual research project and potential secondary use might not be explicated in a project report. This paper outlines two methods for classifying and extracting the impact of publicly funded research projects. The first method is concerned with identifying impact categories and assigning these categories to research projects and their reports by extension by using subject matter experts; not considering the content of research reports. This process resulted in a classification schema that we describe in this paper. With the second method which is still work in progress, impact categories are extracted from the actual text data.

Keywords: research reports, impact assessment, impact categories, corpus analysis

1. Introduction

The disciplinary field and activity of "impact assessment" (IA) are concerned with identifying, estimating, or understanding the consequences of infrastructures, objects, actions, and information on individuals, groups, or society (Latane, 1981). One application domain of IA is scientific research. Research results are mostly available as scientific textual products, e.g., research publications and project reports. It might be challenging for academic institutions, funding organizations[1], and other stakeholders of academic research to reliably identify methods or outcomes mentioned in project reports that have led to additional benefits of the work beyond the project, especially outside of academia. In other words, while academic impact is often achieved through publications and presentations by the researchers who did the work, the impact of research on society might be less obvious and hard to measure. Doing so matters though as the transfer of academic knowledge becomes increasingly important to researchers, funders, and society.

Due to a lack of standardized structure and language use in written descriptions of research projects and results across disciplines, studying and analyzing the impact or impact opportunities of research outcomes requires human domain experts and/or advanced technical solutions to go through the texts and extract the relevant information. For humans, this task is expensive in terms of time and expertise, and automated solutions are yet to be developed. Additionally, manual evaluation is limited by the large and growing number of research papers. As research papers are heterogeneous linguistic products, analyzing them semantically or developing automatic procedures to impact measurement or prediction are challenging and complex tasks.

To address these limitations, in our joint collaborative project TextTransfer[2], we aim to evaluate the impact of publicly funded research projects beyond academia by using an interdisciplinary, mixed-methods approach. The impact evaluation in this project is based on final reports that are collected by the German National Library of Science and Technology (TIB)[3] upon projects completion.

In this paper, we propose a new methodology for capturing and classifying non-academic impact of research projects by combining subject matter expertise with computational techniques (natural language processing, machine learning). We use two methods to identify the impact of research projects: First, we identify external (to the project reports) and objective indicators of impact. Second, we analyze project reports for mentions or indicators of impact. We will compare the results from both methods to better understand the types and magnitude of impact of projects along various dimensions (e.g., monetary versus non-monetary, sociopolitical vs. non-sociopolitical, etc.).

The rest of the paper is organized as follows: in the second section, we synthesize the theoretical foundations of impact studies. In the third section, we explain our two approaches for defining and extracting impact (external versus text-based). In section four, we mention our

[1] Some funding organization have an explicit mission to develop methods for increasing impact and transferring research results.

[2] http://www1.ids-mannheim.de/direktion/fi/projekte/texttransfer.html

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

preliminary outcomes and next steps. Finally, section 5 outlines potential future uses for the project final outcomes.

2. Theory and background on impact

Impact assessment (IA) has been studied and practiced for several decades in various disciplines and application domains, e.g., environmental studies (D. R. Becker, Harris, McLaughlin, & Nielsen, 2003; H. A. Becker, 2001; Vanclay, 2003), psychology (Latane, 1981), and media studies (Nisbet & Aufderheide, 2009; Whiteman, 2004). Across fields, the benefits of IA include facilitating decision-making processes, and minimizing risks while maximizing returns of investments.

The goals with IA are typically to broadly identify and precisely understand a project's future consequences. Gaining a clear and comprehensive understanding of a project is a precondition to be able to achieve these goals. Such an understanding is also key to designing methods for tackling anticipated problems or troubleshooting emerging issues. To gain such an understanding, after proposing a plan or project, scientists consult with domain experts, collect data, and study similar prior projects. As part of theses processes, assessors aim to get familiar with the domain-specific, local, and regional settings, norms, and regulations. Methods for gaining a localized and contextualized understanding of a situation include surveys and interviews. The resulting findings can then inform the planning of proper actions or adjusting plans. In the following sub-sections, we briefly discuss approaches to IA in the fields of environmental studies, information science, and library science.

2.1 Environmental Studies

Vanclay defines "social impact assessment" (SIA) as the study and analysis of the consequences of a planned or unplanned event, the steps that practitioners take to assess the impact of an event, and the development of strategies for monitoring and managing those impacts (Vanclay, 2003). After identifying probable impacts on humans, the economy, or the environment, a plan will be designed and shared with the public. That plan might then be changed due to suggestions and feedback. The updated plan will be delivered to participating organizations. Post-project monitoring may also be conducted (H. A. Becker, 2001; Vanclay, 2006).

SIA was first introduced in the National Environment Policy Act (NEPA) around 1960. Later, scientists formed a committee for "Social Impact Assessment" in order to meet the requirements defined in NEPA for private sector organizations ("Guidelines and principles for social impact assessment," 1995).

The IA approach presented for our study differs from SIA as we conduct assessment ex post facto to identify indicators for or correlations of text-based or project-based features, respectively, with secondary and subsequent (typically after project completion) outcomes of research projects. In the long run, academic research practices could adopt lessons learned from SIA to proactively anticipate lateral or subsequent consequences of their work on society. In fact, some funding agencies require grant applicants to specify the "broader impacts" of their work. For example, the National Science Foundation of the U.S. defines broader impacts as "the potential of the proposed activity - beyond the research, *per se* - to benefit the Nation", which may include promoting education, broadening the "participation of underrepresented groups", enhancing "infrastructure for research and education", advance "scientific and technological understanding", and benefits to society ("Broader impacts review criterion," n.d.).

2.2 Information Science

As IA in environmental studies aims to anticipate potential effects of future actions, IA of media and information focuses on the influence of information on people and society. This perspective has gained attention in recent years as funders and producers aim to measure the impact of information products on people (Diesner, Kim, & Pak, 2014; Diesner & Rezapour, 2015; John & James, 2011; Karlin & Johnson, 2011). A primary goal of information products, producers, and funders is often to raise awareness about issues in the general public (Clark & Abrash, 2011). Data collection and analysis approaches in this area can entail mixed-data and mixed-methods studies, for example, they may combine 1) qualitative analysis of interviews with 2) quantitative analyses of surveys or web metrics.

Impact of information can be divided into influence on the macro, meso, and micro level as explained next.

Macro-level impact refers to changes on the societal level, e.g., legislative and policy changes that result in raised awareness ("Impact glossary," n.d.). Impact of user-generated (e.g., social media) or professionally-generated (e.g., mainstream media) information on society may also entail changes in discourse and culture.

Meso-level impact refers to changes on the corporate and institutional level (Chattoo, 2014; "Impact glossary," n.d.), and can also include change in the structure of communities or the formation of new communities (Chattoo, 2014).

Micro-level impact refers to influence on individual people, such as 1) changes in awareness, 2) affecting behavior, cognition, and emotions, and 3) motivating civic engagement (Barrett & Leddy, 2008; Chattoo, 2014; Clark & Abrash, 2011; Karlin & Johnson, 2011). The aggregate of these effects can also result in the aforementioned higher-level types of impact. Based on surveys, closed group interviews, and data mining techniques, it was found that individuals indicated change in behavior and knowledge associated with watching films (Blakley, Huang, Nahm, & Shin, n.d.; Schiffrin, 2014; Schiffrin & Zuckerman, 2015). Rezapour and Diesner (2017) studied the impact of information products on individuals by analyzing film reviews and identifying and measuring different types of micro-level impact, such as changes versus reaffirmation in personal behavior, cognition, and emotions.

Relating these insights to measuring the impact of research reports, we acknowledge that research outcomes may intend to or potentially have an impact on all of these levels (macro, meso, micro).

2.3 Library Science

IA in the field of library science focuses on designing and creating efficient systems to meet the needs of customers and enhancing customer experience. Additional goals include increasing the influence of libraries, e.g., via outreach activities. To assess library services and systems based on this conceptualization of impact, one needs to 1) understand library or information users and their needs, and 2) employ a combination of qualitative and quantitative research methods (Connaway & Radford, 2016).

In the areas of bibliometrics and scientometrics, the impact of scholarly work on the scientific community has traditionally been measured by considering citation counts, and calculating metrics over these counts, such as the h-index (Bornmann & Daniel, 2005; Hirsch, 2005). More recent efforts, such as the altmetrics movement, also consider the impact on research beyond academic, e.g., by analyzing mentions of research on social and traditional media, or tracking the sharing and reuse of resources and data (Piwowar, 2013; Priem, Taraborelli, Groth, & Neylon, 2010).

The project described in this paper builds upon prior insights from various disciplines as outlined above, and expands alternative ways to study the impact of scholarly work. Our work is based on the assumption that regardless of the type and application domain of impact, information products can affect people, communities or society in a direct or indirect manner. We acknowledge that while the findings from research projects, especially from basic research, may not directly influence people's daily lives, they can lead to fundamental changes and restructurings of different aspects of society in the long run. Making this transition may require transfer from fundamental research to applications, moving from illustrative applications and limited samples to findings of general applicability and scalable performance, etc. Our project aims at identifying text-based indicators for or correlations with such subsequent outcomes.

3. Impact of research projects beyond academia

As mentioned before, the TextTransfer project is concerned with evaluating the impact of research projects based on their final reports. These reports are collected by the TIB after project completion. We do not study the academic impact of these reports or projects, but their impact beyond academia.

Our text corpus for analysis consists of the text version (PDF) of project reports, meta-data about the project (e.g., duration, partners), and meta-data on the reports (e.g., number of pages).

Our work is based on the assumption that reports on projects with subsequent impact beyond academia have text-level characteristics or indicators that can be distinguished from reports of projects with no or little proven effects after completion. For the later intended stage of building a classifier, we also assume that these indicative features not only occur in the reports that we examine, but also generalize to other reports.

Our goal is to develop a computational methodology for detecting and classify impact indicators in large amounts of texts in a short amount of time and with high accuracy. We aim for this work to help libraries and funders to efficiently assess potential future uses of research projects. We hope that this work can also inform efforts to develop automated processes for identifying the potential usages of projects based on scientific texts. Our work is not intended to motivate the reverse engineering of impact (from hopes to texts).

3.1 Dataset

We analyze final reports of publicly funded projects with the specific focus on the question if the results of the projects have been put to usage outside of science after the project ended. A project can have one or more reports; the latter applies for example to projects with multiple independent but collaborating partners. Since the number of reports available in TIB is large[4], we selected a sample based on the following criteria, which all projects in the sample must meet:

- report(s) digitally available in the TIB library (PDFs and metadata),
- project domain: electro-mobility,
- project profile: technology and promotion of innovation,
- project completion: between 2005 and 2015,
- at least two partners,
- at least one academic project partner.

The resulting sample contained about 450 projects. Since the reports are in PDF format, they must be converted into a format suitable for automatic processing. We chose to convert them into both plain text and TEI-I5 format. Since we are only interested in the text content, non-textual data like pictures, complex mathematical typesetting, and table layouts are not being remodeled in the destination formats as doing so is error-prone and of no avail for textual analysis. We acknowledge that these elements might be of use for analysis at some point, but multi-modal data analysis is beyond the scope of the current work on this project.

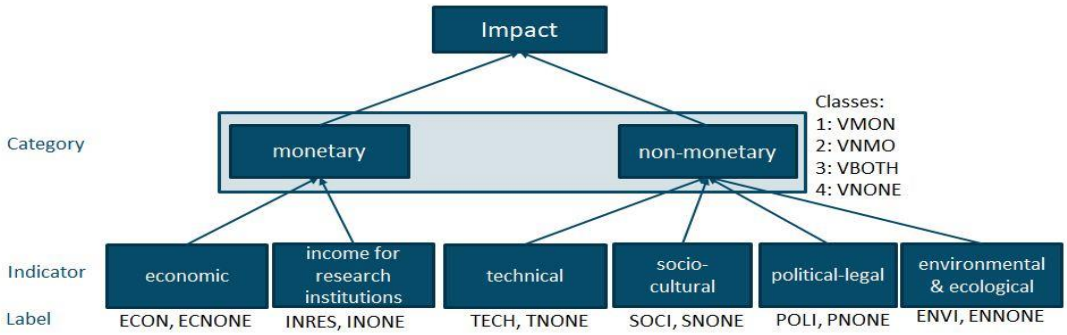
Next, we need to classify the projects with respect to impact on the non-academic community. We do this in two ways: First, by identifying objective evidence of project impact regardless of the reports (detailed in section 3.3). Second, by assessing impact only based on the text content of the project reports (detailed in section 3.4).

3.2 Impact definition and measurement

As mentioned in section 2, IA has been studied and/or practiced for decades in various fields and application domains. For some of these fields and domains, defining impact may be a clear and straightforward task. Also, assessing impact may involve qualitative and quantitative research.

[4] In November 2016 the TIB collection comprised about 256000 printed and 75000 electronic documents of which 65097 are openly accessible in PDF format.

Figure 1: Classification scheme for external impact categories



The process of defining and measuring impact of research projects beyond academia is challenging for the following reasons. The first issue is timing: it might take time after project completion to convert research findings and other outcomes into knowledge, activities, services, products, etc. that affect society. The time span between project completion and impact can vary widely. Also, we are solely relying on project reports of the completed projects. These reports may describe impact that has already been realized (which is easy to identify), or anticipate future impact, which might not be realized (which requires careful distinction between potential and actual impact).

The second issue is defining impact of research projects. Impact can be direct (e.g., a new online service), immediate (within the project lifetime and reporting), indirect (the contribution of the project is not obvious to the public), or delayed (after project completion and reporting). In order to be be able to distinguish these aspects, we use two different approaches to define and measure impact. We will also test the congruence of these approaches.

The first approach is deductive: we define external impact categories of research projects, and let experts assess the impact for every project in our project sample regardless of the project reports. We then perform text analysis techniques to find correlations between the texts and externally defined and identified impact of the related projects. This approach is based on the assumption that some text features in the project reports might correlate with impact categories, which are detailed in section 3.3.

The second approach is to let human coders analyze project reports from our sample of reports, and identify and mark up text-based indicators of impact. We will then use the analysis results for deductive learning. This step is described in section 3.4.

In the final step, we will compare the results from both approaches in order to find out if text-level impact aligns with expert judgment on the project level impact.

3.3 Externally defined impact categories and measurement

Our first approach is to define external impact categories of research projects, and to let experts assign applicable categories to the projects in our sample. In a first step, we defined six objective impact indicators/criteria for research projects in general:

- Economic impact: refers to the use of research results in the private sector, e.g., the development of a business model.
- Income impact: refers to additional income for research institutions, e.g., selling licenses or establishing research contracts.
- Technical impact: refers to technologies that are used outside of the original project, e.g., prototype development or process development.
- Socio-cultural impact: 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: 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: refers to changes of ecological or environmental aspects, e.g. environmental reports or weather data collection.

We then created two higher-order categories that we associated with these six categories: “monetary impact” and “non-monetary impact”, and based on that, four classes: "monetary impact" (VMON), “non-monetary impact” (VNMO), “monetary and non-monetary impact” (VBOTH) and “no impact” (VNONE), see Figure 1.

Monetary impact of a project considers the indicators for economic impact and income impact. The non-monetary impact considers the other four indicators. Any given project must be categorized with one of two possible labels for each of the six indicators: the first label is the positive one (e.g., “ECON” means the project has economic impact), and the second label is the negative one (e.g., “ECNONE” means the project has no economic impact). According to our impact type classification schema (Figure 1), the categories of “monetary impact” and “non-monetary” impact are not collected separately, but derived from the six pre-defined indicators.

In order to label the projects in our sample according to this schema, we tried two methods: First, we did a web search on several projects to find objective evidence for our impact indicators. This process turned out to be heavily time consuming, and not all relevant information about a project's impact could be found online. Therefore, we used a second method: Based on the project reports, the main person per project was identified, contacted via email, the purpose of the contact was explained to them, and they were asked for their permission to perform an interview with them regarding the project. If they agreed, they were asked to answer questions about ten aspects of the project impact. Based on their answers, the project was classified accordingly. For projects with multiple reports, we assigned the impact classes to each report on the project.

3.4 Text-based definition and measurement of impact categories

The impact categories described in the previous section are external to the project reports. While we assume the resulting labels to relate to the project reports, they might be independent of the reports. For this reason, we also pursue a second approach, i.e., identifying impact solely based on the project reports. For this task, we first asked human annotators to read a sample of project reports, and based on that, suggest impact categories that they see in the data. These annotators are not aware of the externally defined classification schema. Hence, the text-based impact categories may or may not overlap with the external ones. In the next step, we will review the suggested categories and synthesize them into a formal system of categories, resulting in a codebook. The codebook will then be used by at least two independent annotators to mark up a larger set of project reports from our sample. After finishing the document annotation and measuring intercoder reliability, we plan to train a classifier for impact types and categories using the annotated data for training, so that we can use the resulting model to label projects automatically for their potential impact.

4. Preliminary outcomes

We have completed the definition of external impact categories, and the project reports in our sample are being labeled accordingly. Identifying the text-based impact categories is work in progress that we will report on in the workshop presentation.

As soon as all projects are labeled using the two different methods, we will extract features, train classifiers, evaluate their accuracy, and conduct an error analysis. Finally, we will test the congruence of the two selected methods for measuring impact.

5. Discussion

Ideally, the outcomes of research projects include or lead to broader impacts, i.e., benefits to society beyond the research project *per se*. With our approach, we hope to allow researchers as well as other stakeholders of publicly funded research to assess how research projects might have different kinds of impact (economic, sociopolitical, environmental, etc.).

By enhancing the meta-data of the project reports with our impact categories, we also want to provide a valuable resource for the interested community. We aim for the approach described in this paper to be applicable to research and application areas beyond electro-mobility. The impact categories might need to be customized for other application domains, but the overall research design should still be applicable.

6. Acknowledgement

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.

7. Bibliographical References

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

Becker, D. R., Harris, C. C., McLaughlin, W. J., & 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.

Blakley, J., Huang, G., Nahm, S., & Shin, H. (n.d.). *Changing appetites & changing minds: Measuring the impact of "Food, Inc."*. Retrieved from Media Impact Project, The USC Annenberg Norman Lear Center.

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

Broader impacts review criterion. (n.d.). Retrieved from https://www.nsf.gov/pubs/2007/nsf07046/nsf07046.jsp, on 03/12/2018.

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

Clark, J., & Abrash, B. (2011). *Social justice documentary: Designing for impact*. Retrieved from Center for Media and Social Impact, School of Communication at American University.

Connaway, L. S., & Radford, M. L. (2016). *Research methods in library and information science*: ABC-CLIO.

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

Diesner, J., & Rezapour, R. (2015). Social computing for impact assessment of social change projects *Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 34-43): Springer.

Guidelines and principles for social impact assessment. (1995). *Environmental Impact Assessment Review, 15*, 11-43.

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

Impact glossary. (n.d.). Retrieved from http://impact.cironline.org/, on 03/12/2018.

John, S., & James, L. (2011). Impact: A practical guide for evaluating community information projects.

Karlin, B., & Johnson, J. (2011). Measuring impact: The importance of evaluation for documentary film campaigns. *M/C Journal, 14*(6).

Latane, B. (1981). The psychology of social impact. *American Psychologist, 36*(4), 343–356.

Nisbet, M. C., & Aufderheide, P. (2009). Documentary film: Towards a research agenda on forms, functions, and impacts. *Mass Communication and Society, 12*(4), 450-456.

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

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto.

Rezapour, R., & 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*, (pp. 1419-1431), ACM, Portland, Oregon, USA.

Schiffrin, A. (2014). *Measuring media impact: Taking the long term view, supporting media independence*. Retrieved from http://www.academia.edu/7033995/Measuring_Media_Impact_Taking_the_long_term_view_supporting_media_independence

Schiffrin, A., & Zuckerman, E. (2015). Can we measure media impact? Surveying the field.

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*, 3-14.

Whiteman, D. (2004). Out of the theaters and into the streets: A coalition model of the political impact of documentary film and video. *Political Communication, 21*, 51-69.