gesis Leibniz Institute for the Social Sciences

Bedeutung von Text Mining am Beispiel der Sozialwissenschaften

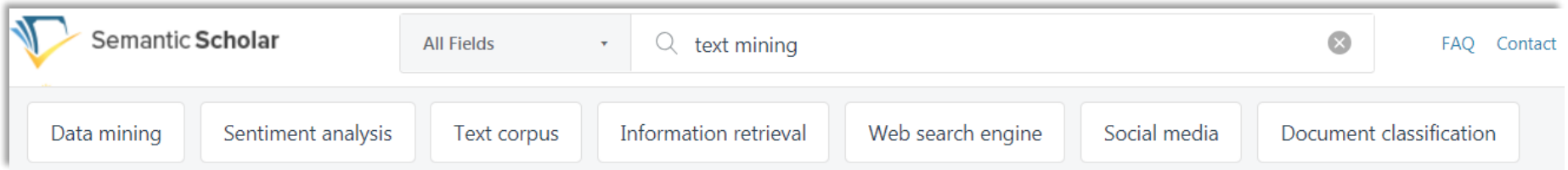
Philipp Mayr

Text und Data Mining (TDM) - Trier

21.06.2018

Member of the Leibniz Association

Text Mining


Semantic Scholar
All Fields
text mining
FAQ
Contact
Data mining
Sentiment analysis
Text corpus
Information retrieval
Web search engine
Social media
Document classification


social networks text mining
social media competitive analysis text mining
visual text mining
text mining tools
market prediction text mining
examining students text mining
pizza industry text mining
study selection activity text mining
systematic reviews text mining
text mining and bibliometrics
text mining techniques
text mining approach
internet studies text mining
live video streaming text mining
online interaction text mining
Gooooooooooogle
1 2 3 4 5 6 7 8 9 10
Weiter

Text Mining in den Sozialwissenschaften

- Politikwissenschaft: Analyse von Wahlkampfreden/-programmen; …
- Wirtschaftswissenschaft: Einfluss von Sozial Media Sentiments auf Börsenentwicklungen; …
- Bibliometrie/Scientometrie: Analyse von einflußreichen Themen innerhalb einer wissenschaftlichen Community; Zitationskont[illegible] Zitationen; …
- …

"Here empirical studies with real users should be performed, closely monitoring the users' actions (involving eyetracking) — see e.g. *[Joachims et al.* 07]."

"In contrast, the work presented in [*Joachims et al. 07*] uses eyetracking for observing users during Web searches, thus registering e.g. the items from the result list users were looking at."

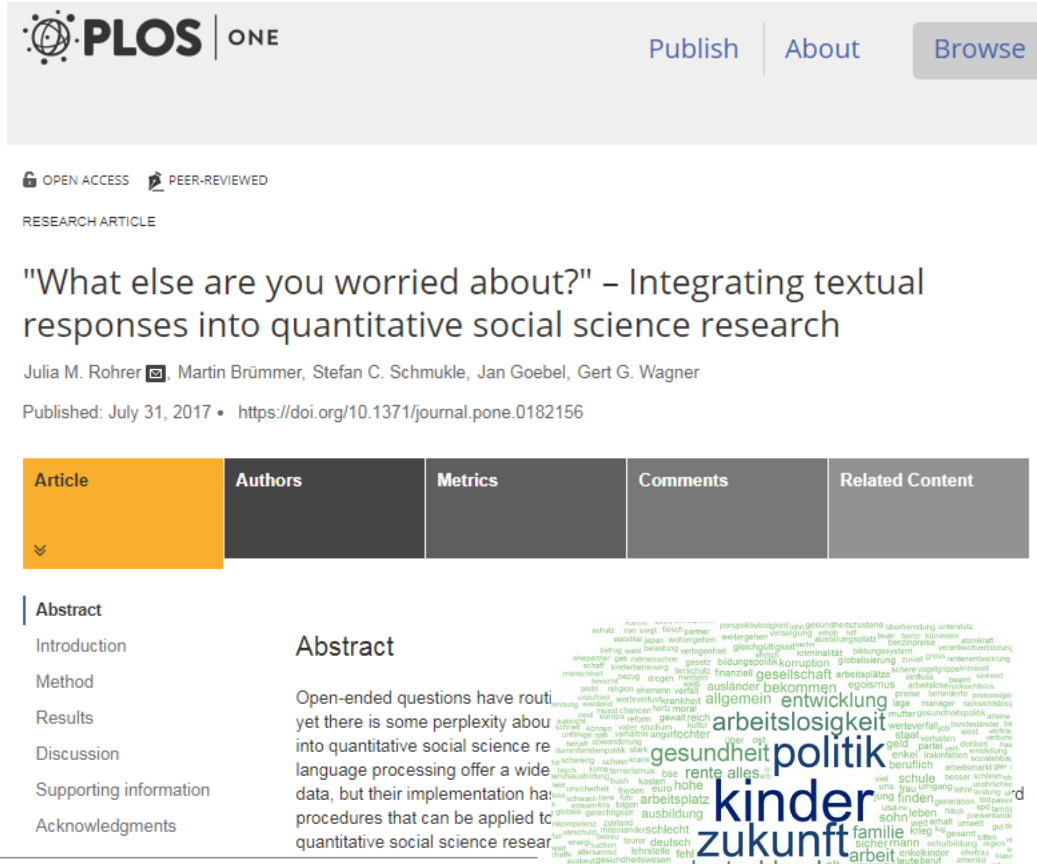
Statement

- Text Mining in den Sozialwissenschaften
  - Starker Anstieg an Text Mining-Aktivitäten
  - Typische autom. Analyseverfahren: Natural Language Processing, Information Retrieval, Named Entity Recognition, Machine Learning, …
  - TM - **Noch kein Standardwerkzeug** der quantitativen Sozialforschung

gesis Leibniz Institute for the Social Sciences

Beispiel 1: Auswertung von offenen Fragen

- Analyse von Open-ended questions in Paneldaten (SOEP)
- Bisherige Analyseverfahren: einfache Keywordsuche, Wortlisten
- Methode: Topic Modelle - data-driven
- Ergebnisse: Erkennung von relevanten Themen und Verbindung der Freitex
  ges
  (sozioc


PLOS ONE | Publish | About | Browse

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

"What else are you worried about?" – Integrating textual responses into quantitative social science research

Julia M. Rohrer, Martin Brümmer, Stefan C. Schmukle, Jan Goebel, Gert G. Wagner

Published: July 31, 2017 • https://doi.org/10.1371/journal.pone.0182156

Article | Authors | Metrics | Comments | Related Content

Abstract
Introduction
Method
Results
Discussion
Supporting information
Acknowledgments

Abstract

Open-ended questions have routi
yet there is some perplexity abou
into quantitative social science re
language processing offer a wide
data, but their implementation has
procedures that can be applied to
quantitative social science resear

Topic	Topic name	Occurrence	Most relevant terms
1	Future of children	19.8%	children, future, occupational, youth, development, worries, grandchild, apprenticeship, son, family
2	Children, youth, school	6.0%	children, youth, school, education, apprenticeship, violence, bad, upbringing, Germany, educational policy
3	Health of family	12.4%	Health, family, man/husband, worries, children, woman/wife, son, parents, mother, daughter
4	Rising prices	6.1%	rising, high, pension, Euro, prices, expensive, costs, unemployment, health reform, taxes
5	Rich and poor	4.0%	social, rich, poor, unjust, Germany, divide, reduction, justice, peace, gap

gesis Leibniz Institute for the Social Sciences

Beispiel 2: Variablen-Erkennung in Volltexten

- **Erkennung von Umfragevariablen in Volltexten**
- Bisherige Analyseverfahren: manuell/intellektuell; Verlinkung von Publikation zu Variablendaten einer Studie
- Methode: Supervised Machine Learning mit Linguistische Analyse
- Ergebnisse: große Trainingsdatensätze für

Mining Social Science Publications for Survey Variables

Andrea Zielinski and **Peter Mutschke**
GESIS - Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
50667 Cologne, Germany
[andrea.zielinski,peter.mutschke]@gesis.org

V39: Belief in life after death o
in Heaven from the ISSP 1998 stud

Abstract

Research in Social Science is usually ... work-in-progress study that seeks to provide a solution to the variable detection task based on supervised machine learning algorithms, using a linguistic analysis pipeline to extract a rich feature set, including terminological concepts and similarity metric scores. Further, we present preliminary results on a small dataset that has been specifically designed for this task, yielding a significant increase in performance over the random baseline.

1 Introduction

In face of the growing number of scientific publications, Text Mining (TM) becomes increasingly important to make hidden knowledge explicit. A particular challenge in this regard is to identify research data citations in scholarly publications, due to their wide variety, ranging from quotations to ... vey question (e.g. *Do you believe in Heaven?*). Therefore, from the perspective of the Social Sciences, having a linkage just to the entire study ... data on the level of variables would have a number of benefits to researchers: It would enable indexing publications by survey variables and discovering publications that discuss the concept of interest (a particular variable). Moreover, it would facilitate a monitoring of the relevance of topical issues (by tracking the use of variables for research) as well as detecting research gaps (by tracking the variables not being addressed by researchers).

The problem, however, is that even though variables are usually assigned a code and a label (e.g. *V39: Belief in life after death* or *V40: Believe in Heaven* from the ISSP 1998 study) as well as the question text from the questionnaire, in practice, authors often do not adhere to citation standards, neither for study names nor for variables. Instead, authors tend to use variations of label and/or question text or combine several variables in one phrase (such as *"...belief in afterlife and Heaven..."* (Neporov and Nepor, 2009)).

Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, 2017

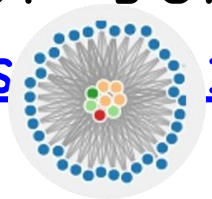
Diskussion

- Warum ist Text Mining noch kein Standardwerkzeug in der Sozialforschung?
  - Verfügbarkeit und Beschaffenheit von Corpora sind ein Problem!
  - Erstellung von Trainingsdatensätzen und Goldstandards z.T. sehr aufwendig!
  - Erweiterte Methodenkenntnisse notwendig!
- Weniger rechtliche Hürden als technisch-methodische Probleme!
  - Publikation von Twitter Tweets lediglich als Tweet-ID Listen

gesis Leibniz Institute for the Social Sciences

Exemplarische Projekte bei GESIS

- OpenMinTeD - Open Mining Infrastructure for Text and Data http://openminted.eu/
- InFoLiS - Integration von Forschungsdaten und Literatur in den Sozialwissenschaften http://infolis.github.io/
- EXCITE - Extraktion von Zitationen aus PDF-Dokumenten http://excite.west.uni-koblenz.de/



EXCITE

References

- Rohrer, J. M., Brümmer, M., Schmukle, S. C., Goebel, J., & Wagner, G. G. (2017). 'What else are you worried about?' - Integrating textual responses into quantitative social science research. PLOS ONE, 12(7), e0182156. https://doi.org/10.1371/journal.pone.0182156
- Zielinski, A., & Mutschke, P. (2017). Mining Social Science Publications for Survey Variables. In Proceedings of the Second Workshop on NLP and Computational Social Science. ACL 2017. http://aclweb.org/anthology/W17-29

Kontakt

Dr. Philipp Mayr

Email: philipp.mayr@gesis.org

Web: http://www.gesis.org

@philipp_mayr

SSOAR

http://ssoar.info

ECIR 2019 Cologne

http://ecir2019.org/

41st European Conference on Information Retrie
in Cologne, Germany, on 14th April to 18th Apr