Kritische Studien zur Demokratie

Gregor Wiedemann

Text Mining for Qualitative Data Analysis in the Social Sciences

A Study on Democratic Discourse in Germany

Springer VS

Kritische Studien zur Demokratie

Die Erforschung demokratischer Praxis aus normativer wie empirischer Perspektive zählt zu den wichtigsten Gegenständen der Politikwissenschaft. Dabei gilt es auch, kritisch Stellung zum Zustand und zu relevanten Entwicklungstrends zeitgenössischer Demokratie zu nehmen. Besonders die Politische Theorie ist Ort des Nachdenkens über die aktuelle Verfasstheit von Demokratie. Die Reihe *Kritische Studien zur Demokratie* versammelt aktuelle Beiträge, die diese Perspektive einnehmen: Getragen von der Sorge um die normative Qualität zeitgenössischer Demokratien versammelt sie Interventionen, die über die gegenwärtige Lage und die künftigen Perspektiven demokratischer Praxis reflektieren. Die einzelnen Beiträge zeichnen sich durch eine methodologisch fundierte Verzahnung von Theorie und Empirie aus.

Gregor Wiedemann

Text Mining for Qualitative Data Analysis in the Social Sciences

A Study on Democratic Discourse in Germany

Springer VS

Gregor Wiedemann
Leipzig, Germany

Dissertation Leipzig University, Germany, 2015

Kritische Studien zur Demokratie
ISBN 978-3-658-15308-3 ISBN 978-3-658-15309-0 (eBook)
DOI 10.1007/978-3-658-15309-0

Library of Congress Control Number: 2016948264

Springer VS

Printed on acid-free paper

This Springer VS imprint is published by Springer Nature
The registered company is Springer Fachmedien Wiesbaden GmbH
The registered company address is: Abraham-Lincoln-Strasse 46, 65189 Wiesbaden, Germany

Preface

Two developments in computational text analysis widen opportunities for qualitative data analysis: amounts of digital text worth investigating are growing rapidly, and progress in algorithmic detection of semantic structures allows for further bridging the gap between qualitative and quantitative approaches. The key factor here is the inclusion of context into computational linguistic models which extends simple word counts towards the extraction of meaning. But, to benefit from the heterogeneous set of text mining applications in the light of social science requirements, there is a demand for a) conceptual integration of consciously selected methods, b) systematic optimization of algorithms and workflows, and c) methodological reflections with respect to conventional empirical research.

This book introduces an integrated workflow of text mining applications to support qualitative data analysis of large scale document collections. Therewith, it strives to contribute to the steadily growing fields of digital humanities and computational social sciences which, after an adventurous and creative coming of age, meanwhile face the challenge to consolidate their methods. I am convinced that the key to success of digitalization in the humanities and social sciences not only lies in innovativeness and advancement of analysis technologies, but also in the ability of their protagonists to catch up with methodological standards of conventional approaches. Unequivocally, this ambitious endeavor requires an interdisciplinary treatment. As a political scientist who also studied computer science with specialization in natural language processing, I hope to contribute to the exciting debate on text mining in empirical research by giving guidance for interested social scientists and computational scientists alike.

Gregor Wiedemann

Contents

List of Figures

List of Tables

List of Abbreviations

AAD	Analyse Automatique du Discours
BMBF	Bundesministerium für Bildung und Forschung
BfV	Bundesamt für Verfassungsschutz
BMI	Bundesministerium des Innern
CA	Content Analysis
CAQDA	Computer Assisted Qualitative Data Analysis
CATA	Computer Assisted Text Analysis
CCA	Computational Content Analysis
CDA	Critical Discourse Analysis
CLARIN	Common Language Resources and Technology Infrastructure
MP	Manifesto Project
CTM	Correlated Topic Model
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DASISH	Digital Services Infrastructure for Social Sciences and Humanities
DH	Digital Humanities
DKP	Deutsche Kommunistische Partei
DTM	Document-Term-Matrix
DVU	Deutsche Volksunion
ESFRI	European Strategic Forum on Research Infrastructures
EU	European Union
FAZ	Frankfurter Allgemeine Zeitung
FdGO	Freiheitlich-demokratische Grundordnung
FQS	Forum Qualitative Social Research
FRG	Federal Republic of Germany
GDR	German Democratic Republic

GTM	Grounded Theory Methodology
IDF	Inverse Document Frequency
IR	Information Retrieval
JSD	Jensen–Shannon Divergence
KPD	Kommunistische Partei Deutschlands
KWIC	Key Word in Context
LDA	Latent Dirichlet Allocation
LL	Log-likelihood
LSA	Latent Semantic Analysis
ML	Machine Learning
MAXENT	Maximum Entropy
MAP	Mean Average Precision
MDS	Multi Dimensional Scaling
MWU	Multi Word Unit
NB	Naive Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NPD	Nationaldemokratische Partei Deutschlands
NSDAP	Nationalsozialistische Deutsche Arbeiterpartei
OCR	Optical Character Recognition
OLAP	Online Analytical Processing
ORC	Open Research Computing
OWL	Web Ontology Language
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PDS	Partei des Demokratischen Sozialismus
PMI	Pointwise Mutual Information
POS	Part of Speech
QCA	Qualitative Content Analysis
QDA	Qualitative Data Analysis
RAF	Rote Armee Fraktion
RDF	Resource Description Framework
RE	Requirements Engineering
REP	Die Republikaner
RMSD	Root Mean-Square Deviation

SE	Software Engineering
SECG	Semantically Enriched Co-occurrence Graph
SED	Sozialistische Einheitspartei Deutschlands
SOM	Self Organizing Map
SPD	Sozialdemokratische Partei Deutschlands
SRP	Sozialistische Reichspartei
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TM	Text Mining
TTM	Term-Term-Matrix
UN	United Nations
VSM	Vector Space Model
WASG	Wahlalternative Arbeit und Soziale Gerechtigkeit
XML	Extensible Markup Language

1. Introduction: Qualitative Data Analysis in a Digital World

Digitalization and informatization of science during the last decades have widely transformed the ways in which empirical research is conducted in various disciplines. Computer-assisted data collection and analysis procedures even led to the emergence of new subdisciplines such as bioinformatics or medical informatics. The humanities (including social sciences)[1] so far seem to lag somewhat behind this development—at least when it comes to analysis of textual data. This is surprising, considering the fact that text is one of the most frequently investigated data types in philologies as well as in social sciences like sociology or political science. Recently, there have been indicators that the digital era is constantly gaining ground also in the humanities. In 2009, fifteen social scientists wrote in a manifesto-like article in the journal "Science":

> "The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven 'computational social science' has been much slower. [...] But computational social science is occurring – in internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency" (Lazer et al., 2009, p. 721).

In order not to leave the field to private companies or governmental agencies solely, they appealed to social scientists to further embrace computational technologies. For some years, developments marked by

[1] In the German research tradition the disciplines of social sciences and other disciplines of the humanities are separated more strictly (*Sozial- und Geisteswissenschaften*). Thus, I hereby emphasize that I include social sciences when referring to the (digital) humanities.

G. Wiedemann, *Text Mining for Qualitative Data Analysis in the Social Sciences*,
Kritische Studien zur Demokratie, DOI 10.1007/978-3-658-15309-0_1

popular buzzwords such as *digital humanities*, *big data* and *text and data mining* blaze the trail through the classical publications. Within the humanities, social sciences appear as pioneers in application of these technologies because they seem to have a 'natural' interest for analyzing semantics in large amounts of textual data, which firstly is nowadays available and secondly rises hope for another type of representative studies beyond survey research. On the other hand, there are well established procedures of manual text analysis in the social sciences which seem to have certain theoretical or methodological prejudices against computer-assisted approaches of large scale text analysis. The aim of this book is to explore ways of systematic utilization of (semi-)automatic computer-assisted text analysis for a specific political science research question and to evaluate on its potential for integration with established manual methods of qualitative data analysis. How this is approached will be clarified further in Section 1.4 after some introductory remarks on digital humanities and its relation to social sciences.

But first of all, I give two brief definitions on the main terms in the title to clarify their usage throughout the entire work. With *Qualitative Data Analysis* (QDA), I refer to a set of established procedures for analysis of textual data in social sciences—e.g. Frame Analysis, Grounded Theory Methodology, (Critical) Discourse Analysis or (Qualitative) Content Analysis. While these procedures mostly differ in underlying theoretical and methodological assumptions of their applicability, they share common tasks of analysis in their practical application. As Schönfelder (2011) states, "qualitative analysis at its very core can be condensed to a close and repeated review of data, categorizing, interpreting and writing" (§ 29). Conventionally, this process of knowledge extraction from text is achieved by human readers rather intuitively. QDA methods provide systematization for the process of structuring information by identifying and collecting relevant textual fragments and assigning them to newly created or predefined semantic concepts in a specific field of knowledge. The second main term *Text Mining* (TM) is defined by Heyer (2009, p. 2) as a set of "computer based methods for a semantic analysis of text that help

to automatically, or semi-automatically, structure text, particularly very large amounts of text". Interestingly, this definition comprises of some analogy to procedures of QDA with respect to structure identification by repeated data exploration and categorization. While manual and (semi-)automatic methods of structure identification differ largely with respect to certain aspects, the hypothesis of this study is that the former may truly benefit from the latter if both are integrated in a well-specified methodological framework. Following this assumption, I strive for developing such a framework to answer the question

1. How can the application of (semi-)automatic TM services support qualitative text analysis in the social sciences, and

2. extend it with a quantitative perspective on semantic structures towards a mixed method approach?

1.1. The Emergence of "Digital Humanities"

Although computer assisted content analysis already has a long tradition, so far it did not prevail as a widely accepted method within the QDA community. Since computer technology became widely available at universities during the second half of the last century, social science and humanities researchers have used it for analyzing vast amounts of textual data. Surprisingly, after 60 years of experience with computer-assisted automatic text analysis and a tremendous development in information technology, it still is an uncommon approach in the social sciences. The following section highlights two recent developments which may change the way qualitative data analysis in social sciences is performed: firstly, the rapid growth of the availability of digital text worth to investigate and, secondly, the improvement of (semi-)automatic text analysis technologies which allows for further bridging the gap between qualitative and quantitative text analysis. In consequence, the use of text mining cannot be characterized only as a further development of traditional quantitative content analysis beyond communication and media studies. Instead, computational

linguistic models aiming towards the extraction of meaning comprise opportunities for the coalescence of former opposed research paradigms in new mixed method large-scale text analyses.

Nowadays, Computer Assisted Text Analysis (CATA) means much more than just counting words.[2] In particular, the combination of pattern-based and complex statistical approaches may be applied to support established qualitative data analysis designs and open them up to a quantitative perspective (Wiedemann, 2013). Only a few years ago, social scientists somewhat hesitantly started to explore its opportunities for their research interest. But still, social science truly has much unlocked potential for applying recently developed approaches to the myriads of digital texts available these days. Chapter 2 introduces an attempt to systematize the existing approaches of CATA from the perspective of a qualitative researcher. The suggested typology is based not only on the capabilities contemporary computer algorithms provide, but also on their notion of context. The perception of context is essential in a two-fold manner: From a qualitative researcher's perspective, it forms the basis for what may be referred to as meaning; and from the Natural Language Processing (NLP) perspective it is the decisive source to overcome the simple counting of character strings towards more complex models of human language and cognition. Hence, the way of dealing with context in analysis may act as decisive bridge between qualitative and quantitative research designs.

Interestingly, the quantitative perspective on qualitative data is anything but new. Technically open-minded scholars more than half a century ago initiated a development using computer technology for textual analysis. One of the early starters was the Italian theologist Roberto Busa, who became famous as "pioneer of the digital humanities" for his project "Index Thomasticus" (Bonzio, 2011). Started in 1949—with a sponsorship by IBM—this project digitalized and indexed the complete work of Thomas Aquinas and made it publicly

[2] In the following, I refer to CATA as the complete set of software-based approaches of text analysis, not just Text Mining.

available for further research (Busa, 2004). Another milestone was the software THE GENERAL INQUIRER, developed in the 1960s by communication scientists for the purpose of computer-assisted content analysis of newspapers (Stone et al., 1966). It made use of frequency counts of keyword sets to classify documents into given categories. But, due to a lack of theoretical foundation and exclusive commitment to deductive research designs, emerging qualitative social research remained skeptical about those computer-assisted methods for a long time (Kelle, 2008, p. 486). It took until the late 1980s, when personal computers entered the desktops of qualitative researchers, that the first programs for supporting qualitative text analysis were created (Fielding and Lee, 1998). Since then, a growing variety of software packages, like MAXQDA, ATLAS.ti or NVivo, with relatively sophisticated functionalities, became available, which make life much easier for qualitative text analysts. Nonetheless, the majority of these software packages has remained "truly qualitative" for a long time by just replicating manual research procedures of coding and memo writing formerly conducted with pens, highlighters, scissors and glue (Kuckartz, 2007, p. 16).

This once justified methodological skepticism against computational analysis of qualitative data might be one reason for qualitative social research lagging behind in a recent development labeled by the popular catchword Digital Humanities (DH) or 'eHumanities'. In contrast to DH, which was established at the beginning of the 21st century (Schreibman et al., 2004), the latter term emphasizes the opportunities of computer technology not only for digitalization, storage and management of data, but also for analysis of (big) data repositories.[3]

Since then, the digitalization of the humanities has grown in big steps. Annual conferences are held, institutes and centers for DH are founded and new professorial chairs have been set up. In 2006, a group

[3] A third term, "computational humanities", is suggested by Manovich (2012). It emphasizes the fact that additionally to the digitalized version of classic data of the humanities, new forms of data emerge by connection and linkage of data sources. This may apply to 'retro-digitalized' historic data as well as to 'natively digital' data in the worldwide communication of the 'Web 2.0'.

of European computer linguists developed the idea for a long-term project related to all aspects of language data research leading to the foundation of the *Common Language Resources and Technology Infrastructure* (CLARIN)[4] as part of the *European Strategic Forum on Research Infrastructures* (ESFRI). CLARIN is planned to be funded with 165 million Euros over a period of 10 years to leverage digital language resources and corresponding analysis technologies. Interestingly, although mission statements of the transnational project and its national counterparts (for Germany CLARIN-D) speak of humanities *and* social sciences as their target groups[5], few social scientists have engaged in the project so far. Instead, user communities of philologists, anthropologists, historians and, of course, linguists are dominating the process. In Germany, for example, a working group for social sciences in CLARIN-D concerned with aspects of computational content analysis was founded not before late 2014. This is surprising, given the fact that textual data is one major form of empirical data many qualitatively-oriented social scientists use. Qualitative researchers so far seem to play a minor role in the ESFRI initiatives. The absence of social sciences in CLARIN is mirrored in another European infrastructure project as well: the *Digital Research Infrastructure for the Arts and Humanities* (DARIAH)[6] focuses on data acquisition, research networks and teaching projects for the Digital Humanities, but does not address social sciences directly. An explicit QDA perspective on textual data in the ESFRI context is only addressed in the *Digital Services Infrastructure for Social Sciences and Humanities* (DASISH).[7] The project perceives digital "qualitative social science data", i.e. "all non-numeric data in order to answer specific research questions" (Gray, 2013, p. 3), as subject for quality assurance, archiving and accessibility. Qualitative researchers in the DASISH context acknowledge that "the inclusion of qualitative data represents

[4] http://clarin.eu

[5] "CLARIN-D: a web and centres-based research infrastructure for the social sciences and humanities" (http://de.clarin.eu/en/home-en.html).

[6] http://dariah.eu

[7] http://dasish.eu

an important opportunity in the context of DASISH's focus on the development of interdisciplinary 'cross-walks' between the humanities and social sciences" reaching out to "quantitative social science", while at the same time highlighting their "own distinctive conventions and traditions" (ibid., p. 11) and largely ignoring opportunities for computational analysis of digitized text.

Given this situation, why has social science reacted so hesitantly to the DH development and does the emergence of 'computational social science' compensate for this late-coming? The branch of qualitative social research devoted to understanding instead of explaining avoided mass data—reasonable in the light of its self-conception as a counterpart to the positivist-quantitative paradigm and scarce analysis resources. But, it left a widening gap since the availability of digital textual data, algorithmic complexity and computational capacity has been growing exponentially during the last decades. Two humanist scholars highlighted this development in their recent work. Since 2000, the Italian literary scholar Franco Moretti has promoted the idea of "distant reading." To study actual world literature, which he argues is more than the typical Western canon of some hundred novels, one cannot "close read" all books of interest. Instead, he suggests making use of statistical analysis and graphical visualizations of hundreds of thousands of texts to compare styles and topics from different languages and parts of the world (Moretti, 2000, 2007). Referring to the Google Books Library Project the American classical philologist Gregory Crane asked in a famous journal article: "What do you do with a Million Books?" (2006). As possible answer he describes three fundamental applications: digitalization, machine translation and information extraction to make the information buried in dusty library shelves available to a broader audience. So, how should social scientists respond to these developments?

1.2. Digital Text and Social Science Research

It is obvious that the growing amount of digital text is of special interest for the social sciences as well. There is not only an ongoing stream of online published newspaper articles, but also corresponding user discussions, internet forums, blogs and microblogs as well as social networks. Altogether, they generate tremendous amounts of text impossible to close read, but worth further investigation. Yet, not only current and future social developments are captured by 'natively' digital texts. Libraries and publishers worldwide spend a lot of effort retro-digitalizing printed copies of handwritings, newspapers, journals and books. The project Chronicling America by the Library of Congress, for example, scanned and OCR-ed[8] more than one million pages of American newspapers between 1836 and 1922. The Digital Public Library of America strives for making digitally available millions of items like photographs, manuscripts or books from numerous American libraries, archives and museums. Full-text searchable archives of parliamentary protocols and file collections of governmental institutions are compiled by initiatives concerned with open data and freedom of information. Another valuable source, which will be used during this work, are newspapers. German newspaper publishers like the *Frankfurter Allgemeine Zeitung*, *Die Zeit* or *Der Spiegel* made all of their volumes published since their founding digitally available (see Table 1.1). Historical German newspapers of the former German Democratic Republic (GDR) also have been retro-digitized for historical research.[9]

Interesting as this data may be for social scientists, it becomes clear that single researchers cannot read through all of these materials. Sampling data requires a fair amount of previous knowledge on the topics of interest, which makes especially projects targeted to a long investigation time frame prone to bias. Further, it hardly enables

[8] Optical Character Recognition (OCR) is a technique for the conversion of scanned images of printed text or handwritings into machine-readable character strings.

[9] http://zefys.staatsbibliothek-berlin.de/ddr-presse

Table 1.1.: Completely (retro-)digitized long term archives of German newspapers.

Publication	Digitized volumes from
Die Zeit	1946
Hamburger Abendblatt	1948
Der Spiegel	1949
Frankfurter Allgemeine Zeitung	1949
Bild (Bund)	1953
Tageszeitung (taz)	1986
Süddeutsche Zeitung	1992
Berliner Zeitung	1945–1993
Neue Zeit	1945–1994
Neues Deutschland	1946–1990

researchers to reveal knowledge structures on a collection-wide level in multi-faceted views as every sample can only lead to inference on the specific base population the sample was drawn from. Technologies and methodologies supporting researchers to cope with these mass data problems become increasingly important. This is also one outcome of the KWALON Experiment the journal Forum Qualitative Social Research (FQS) conducted in April 2010. For this experiment, different developer teams of software for QDA were asked to answer the same research questions by analyzing a given corpus of more than one hundred documents from 2008 and 2009 on the financial crisis (e.g. newspaper articles and blog posts) with their product (Evers et al., 2011). Only one team was able to include all the textual data in its analysis (Lejeune, 2011), because they did not use an approach replicating manual steps of qualitative analysis methods. Instead, they implemented a semi-automatic tool which combined the automatic retrieval of key words within the text corpus with a supervised, data-driven dictionary learning process. In an iterated coding process, they "manually" annotated text snippets suggested

by the computer, and they simultaneously trained a (rather simple) retrieval algorithm generating new suggestions. This procedure of "active learning" enabled them to process much more data than all other teams, making pre-selections on the corpus unnecessary. However, according to their own assessment they only conducted a more or less exploratory analysis which was not able to dig deep into the data. Nonetheless, while Lejeune's approach points into the targeted direction, the present study focuses on exploitation of more sophisticated algorithms for the investigation of collections from hundreds up to hundreds of thousands of documents.

The potential of TM for analyzing big document collections has been acknowledged in 2011 by the German government as well. In a large funding line of the German Federal Ministry of Education and Research (BMBF), 24 interdisciplinary projects in the field of eHumanities were funded for three years. Research questions of the humanities and social science should be approached in joint cooperation with computer scientists. Six out of the 24 projects have a dedicated social science background, thus fulfilling the requirement of the funding line which explicitly had called qualitatively researching social scientists for participation (BMBF, 2011).[10] With their methodological focus on eHumanities, all these projects do not strive for standardized application of generic software to answer their research questions. Instead, each has to develop its own way of proceeding, as

[10] *Analysis of Discourses in Social Media* (http://www.social-media-analytics.org); *ARGUMENTUM* – Towards computer-supported analysis, retrieval and synthesis of argumentation structures in humanities using the example of jurisprudence (http://argumentum.eear.eu); *eIdentity* – Multiple collective identities in international debates on war and peace (http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html); *ePol* – Post-democracy and neoliberalism. On the usage of neoliberal argumentation in German federal politics between 1949 and 2011 (http://www.epol-projekt.de); *reSozIT* – "Gute Arbeit" nach dem Boom. Pilotprojekt zur Längsschnittanalyse arbeitssoziologischer Betriebsfallstudien mit neuen e-Humanities-Werkzeugen (http://www.sofi-goettingen.de/index.php?id=1086); *VisArgue* – Why and when do arguments win? An analysis and visualization of political negotiations (http://visargue.uni-konstanz.de)

well as to reinvent or adapt existing analysis technologies for their specific purpose. For the moment, I assume that generic software for textual analysis usually is not appropriate to satisfy specific and complex research needs. Thus, paving the way for new methods requires a certain amount of willingness to understand TM technologies together with open-mindedness for experimental solutions from the social science perspective. Ongoing experience with such approaches may lead to best practices, standardized tools and quality assurance criteria in the nearby future. To this end, this book strives to make some worthwhile contribution to the extension of the method toolbox of empirical social research. It was realized within and largely profited from the eHumanities-project *ePol – Post-democracy and Neoliberalism* which investigated aspects of qualitative changes of the democracy in the Federal Republic of Germany (FRG) using TM applications on large newspaper collections covering more than six decades of public media discourse (Wiedemann et al., 2013; Lemke et al., 2015).

1.3. Example Study: Research Question and Data Set

The integration of QDA with methods of TM is developed against the background of an exemplary study concerned with longitudinal aspects of democratic developments in Germany. The political science research question investigated for this study deals with the subject of "democratic demarcation". Patterns and changes of patterns within the public discourse on this topic are investigated with TM applications over a time period of several decades. To introduce the subject, I first clarify what "democratic demarcation" refers to. Then, I introduce the data set on which the investigation is performed.

1.3.1. Democratic Demarcation

Democratic political regimes have to deal with a paradox circumstance. On the one hand, the democratic ideal is directed to allow as much freedom of political participation as possible. On the other hand, this freedom has to be defended against political ideas, activities or groups who strive for abolition of democratic rights of participation. Consequently, democratic societies dispute on rules to decide which political actors and ideas take legitimate positions to act in political processes and democratic institutions and, vice versa, which ideas, activities or actors must be considered as a threat to democracy. Once identified as such, opponents of democracy can be subject to oppressive countermeasures by state actors such as governmental administrations or security authorities interfering in certain civil rights. Constitutional law experts as well as political theorists point to the fact that these measures may yield towards undemocratic qualities of the democratic regime itself (Fisahn, 2009; Buck, 2011). Employing various TM methods in an integrated manner on large amounts of news articles from public media this study strives for revealing how democratic demarcation was performed in Germany over the past six decades.

1.3.2. Data Set

The study is conducted on a data set consisting of newspaper articles of two German premium newspapers – the weekly newspaper *Die Zeit* and the daily newspaper *Frankfurter Allgemeine Zeitung (FAZ)*. The *Die Zeit* collection comprises of the complete (retro-)digitized archive of the publication from its foundation in 1946 up to 2011. But, as this study is concerned with the time frame of the FRG founded on May 23rd 1949, I skip all articles published before 1950. The FAZ collection comprises of a representative sample of all articles published between 1959 and 2011.[11] The FAZ sample set was drawn from the

[11] The newspaper data was obtained directly from the publishers to be used in the ePol-project (see Section 1.2). The publishers delivered Extensible Markup Language (XML) files which contained raw texts as well as meta data for

Table 1.2.: Data set for the example study on democratic demarcation.

Publication	Time period	Issues	Articles	Size
Die Zeit	1950–2011	3,398	384,479	4.5 GB
FAZ	1959–2011	15,318	200,398	1.1 GB

complete data set of all articles published during the aforementioned time period by the following procedure:

1. select all articles of category "Meinung" (op-ed commentaries) published in the sections "Politik" (politics), "Wirtschaft" (economics) and "Feuilleton" (feature) and put them into the sample set; then
2. select all articles published in the sections "Politik", "Wirtschaft" and "Feuilleton"
   - which do not belong to the categories "Meinung" or "Rezension" (review),
   - order them by date, and
   - put every twelfth article of this ordered list into the sample set.

The strategy applied to the FAZ data selects about 15 percent of all articles published in the three newspaper sections taken into account. It guarantees that there are only sections included in the sample set which are considered as relevant, and that there are many articles expressing opinions and political positions. Furthermore, it also ensures that the distribution of selected articles over time is directly proportional to the distribution of articles in the base population. Consequently, distributions of language use in the sample can be regarded as representative for all FAZ articles in the given sections over the entire study period.

each article. Meta data comprises of publishing date, headline, subheading, paragraphs, page number, section and in some cases author names.

1.4. Contributions and Structure of the Study

Computer algorithms of textual analysis do not understand texts in a way humans do. Instead they model meaning by retrieving patterns, counting of events and computation of latent variables indicating certain aspects of semantics. The better these patterns overlap with categories of interest expressed by human analysts, the more useful they are to support conventional QDA procedures. Thus, to exploit benefits from TM in the light of requirements from the social science perspective, there is a demand for

1. conceptual integration of consciously selected methods to accomplish analysis specific research goals,

2. systematic adaptation, optimization and evaluation of workflows and algorithms, and

3. methodological reflections with respect to debates on empirical social research.

On the way to satisfy these demands, this introduction has already shortly addressed the interdisciplinary background concerning the *digitalization* of the humanities and its challenges and opportunities for the social sciences. In Chapter 2, methodological aspects regarding qualitative and quantitative research paradigms are introduced to sketch the present state of CATA together with new opportunities for content analysis. In Section 2.2 of this chapter technological foundations of the application of text mining are introduced briefly. Specifically, it covers aspects of representation of semantics in computational text analysis and introduces approaches of (pre-)processing of textual data useful for QDA. Section 2.3 introduces exemplary applications in social science studies. Beyond that, it suggests a new typology of these approaches regarding their notion of context information. This aims to clarify why nowadays TM procedures may be much more compatible with manual QDA methods than earlier approaches such as computer assisted keyword counts dating back to the 1960s have been.

Chapter 3 introduces an integrated workflow of specifically adapted text mining procedures to support conventional qualitative data analysis. It makes a suggestion for a concrete analysis process chain to extract information from a large collection of texts relevant for a specific social science research question. Several technologies are adapted and combined to approach three distinctive goals:

1. *Retrieval of relevant documents:* QDA analysts usually are faced with the challenge to identify document sets from large base populations relevant for rather abstract research questions which cannot be described by single keywords alone. Section 3.1 introduces an Information Retrieval (IR) approach for this demand.

2. *Inductive exploration of collections:* Retrieved collections of (potentially) relevant documents are still by far too large to be read closely. Hence, Section 3.2 provides exploratory tools which are needed to extract meaningful structures for 'distant reading' and good (representative) examples of semantic units for qualitative checks to fruitfully integrate micro- and macro-perspectives on the research subject.

3. *(Semi-)automatic coding:* For QDA categories of content usually are assigned manually to documents or parts of documents. Supervised classification in an active learning scenario introduced in Section 3.3 allows for algorithmic classification of large collections to validly measure category proportions and trends. It especially deals with the considerably hard conditions for machine learning in QDA scenarios.

Technologies used in this workflow are optimized and, if necessary, developed further with respect to requirements from the social science perspective. Among other things, applied procedures are

- key term extraction for dictionary creation,
- document retrieval for selection of sub-corpora,
- thematic and temporal clustering via topic models,