Bremen, March 2025

Armin Müller

Text Mining and Document Classification Workflows for Chinese Administrative Documents

WeSIS — Technical papers
No. 17

Global Dynamics of Social Policy CRC 1342

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Armin Müller
Text Mining and Document Classification Workflows for Chinese Administrative Documents
SFB 1342 Technical Paper Series, 17
Bremen: SFB 1342, 2025

University of Bremen

SFB 1342 Globale Entwicklungsdynamiken von Sozialpolitik /
CRC 1342 Global Dynamics of Social Policy

Postadresse / Postaddress:
Postfach 33 04 40, D - 28334 Bremen

Website:
https://www.socialpolicydynamics.de

[DOI https://doi.org/10.26092/elib/3755]
[ISSN 2700-0389]

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG)
Projektnummer 374666841 – SFB 1342

Armin Müller

Text Mining and Document Classification Workflows for Chinese Administrative Documents

SFB 1342
No. 17

Text Mining and Document Classification Workflows for Chinese Administrative Documents

Armin Müller

Index

About the author

Armin Müller: CRC 1342, University of Bremen, armmueller@constructor.university

Abstract

Background: The political system of the People's Republic of China features a combination of political centralization and administrative decentralization, which makes it one of the most decentralized political systems in the world. The case of social insurance is illustrative of this phenomenon: the national level enacts general laws and regulations, which are further specified at the first sub-national level – by governments at provincial level. But social insurance systems like health insurance or unemployment insurance are typically organized at the second or third sub-national level. Government and administration of prefectural cities and counties pool the funds within their jurisdictions, and enact regulations that ultimately determine inclusiveness and the scope of benefits.

Aim: The aim of this paper is to present approaches to reconstruct the regulatory differences at sub-national level, and to leverage the results for quantitative and qualitative analysis. It provides an introduction to the ongoing document analysis work in project B05 of the CRC 1342 in Bremen. Furthermore, it enables researchers in social-scientific China studies to sort large amounts of regulatory documents by relevance, and to connect regulatory data to survey data or sub-national time series.

Content: This technical paper presents step-by-step the creation of a database to organize the documents, and two workflows to extract information for qualitative and quantitative analysis. The two workflows presented do not exhaust the possibilities of the approach, but merely provide examples used in ongoing publication projects. A complementary GitHub repository provides the code files needed for implementation.

Complementary GitHub repository: https://github.com/arminmueller81/health_insurance_coverage

Key words: Text as data, text classification, machine learning, neural networks, China, administrative documents, legislation

1. Introduction

This technical paper describes workflows for text mining and document classification with large amounts of regulatory documents from the People's Republic of China (PRC) for quantitative and qualitative analysis. It focuses on social insurance systems in healthcare and unemployment as case studies and walks the reader through the main steps of the processes. The paper complements ongoing research in the Collaborative Research Centre (CRC) 1342 "Global Dynamics of Social Policy," and provides a guideline for social scientists working on China. A complementary GitHub repository provides code files in R and Python for sections 3 and 5 (https://github.com/arminmueller81/health_insurance_coverage).

Tools and skill requirements: Basic familiarity with MS-Access and Structured Query Language (SQL) is an indispensable requirement. They offer powerful tools that suffice for many data analysis scenarios, in which keywords and combinations of keywords are sufficient for identifying relevant documents and sentences. Some basic coding and text mining skills (especially: regular expressions) are required for initial data cleaning and organization, which could be accomplished either in R or in Python. Machine learning tools are helpful in answering more complex questions regarding the content, which go beyond simple keyword searches. While conventional algorithms are available in both R and Python, some of the more advanced Natural Language Processing algorithms are only available in Python, and the complementary repository provides Python code for these tasks.

Why is it important? Regulatory documents in the PRC are crucial for policy analysis due to particularities in – first – the structure of the polity and – second – the policy process. First, the PRC has a multilevel system of government characterized by political centralization and administrative decentralization.[1] For social insurance, this means that the central government enacts laws and regulations for a relatively uniform set of institutions to be implemented nationwide. Governments at the first sub-national level – the provincial level – usually specify the guidelines from the central level. Social insurance funds for formal urban employees are usually managed at the second sub-national level – the prefectural level. Funds for the residency-based insurance systems are often still managed at the third sub-national level – the county level. Laws are typically enacted at central level, whereas the lower levels rely on administrative decrees of various types in their operations. In a country of continental dimensions, there is substantial sub-national variation in policy implementation, which is reflected in provincial and local decrees rather than in laws.

Second, the policy process in the PRC tends to rely on extensive experimentation during the stages of agenda setting, policy formulation and implementation (Heilmann 2008). Central legislation tends to conclude the implementation, rather than mark the transition from policy formulation to implementation as in OECD countries. In the case of social insurance, the PRC transformed the old planned-economy system into modern social insurance systems for urban formal employees during the 1980s and 1990s (Duckett 2011; Frazier 2010; Liu 2015; Müller and ten Brink 2022; Solinger 2005). In the 2000s, the government furthermore created health and pension insurance systems for the remainder of the population not working in the formal sector (Müller 2016; 2017). By contrast, the Social Insurance Law (*Shehui baoxian fa*) was only enacted by the National People's Congress in 2010, when all the programs were already rolled out nationwide. What's more, the specifications of the law remained vague in many respects. Also, some specifications of the law were already outdated a few years after its enactment.[2] In sum, material policy development is primarily codified by administrative documents of a less legally

1 The public finance system facilitates fiscal imbalances by concentrating fiscal revenues at the higher levels, and fiscal expenditures at the lower levels (Wong 2009; World Bank 2002).

2 Most notably, the urban and rural pension systems for residents were integrated in the years following the enactment of the social insurance law. In recent years, the government furthermore initiated the integration of health and maternity insurance for formal employees.

binding character than laws due to structural aspects of the polity, and due to idiosyncrasies of legal culture and the policy process.

How is the article structured? Section 2 explains how the administrative documents were procured. Section 3 discusses how a database was created from a large number of individual documents. The information was stored in two formats: first as full documents, and second as individual sentences. Section 4 introduces a workflow based on SQL queries used in a forthcoming publication. Section 5 presents a text classification workflow based on machine learning. Section 6 provides a summary of the approach and its potential.

2. Data collection

Administrative documents from the PRC can be accessed via an online database named PKUlaw (www.pkulaw.com), which is operated by Beijing University. It can be accessed via the CrossAsia platform (www.crossasia.org) from Germany as *Beida Falü Xinxiwang* (北大法律信息网), or from various Universities in China and the Asia-Pacific region. The interface offers separate access to central laws and regulations (*zhongyang fagui*), and sub-national ones (*difang fagui*). Screenshot 1 below shows a search for documents with the term "unemployment insurance" (*shiye baoxian*) in the title on November 12, 2024. The latter were mostly issued by provincial- and prefectural-level organs. There are additional filters for several categories, including the enacting organization (*zhiding jiguan*), or the year of enactment, and – for sub-national documents – the province.

The database has comprehensive coverage of documents at central and provincial level. The prefectural level is overall rather comprehensive, but for certain topics, there may be gaps due to either lack of regulatory activity at that level or failure to make the regulatory documents available. Coverage of county-level jurisdictions is more sporadic.

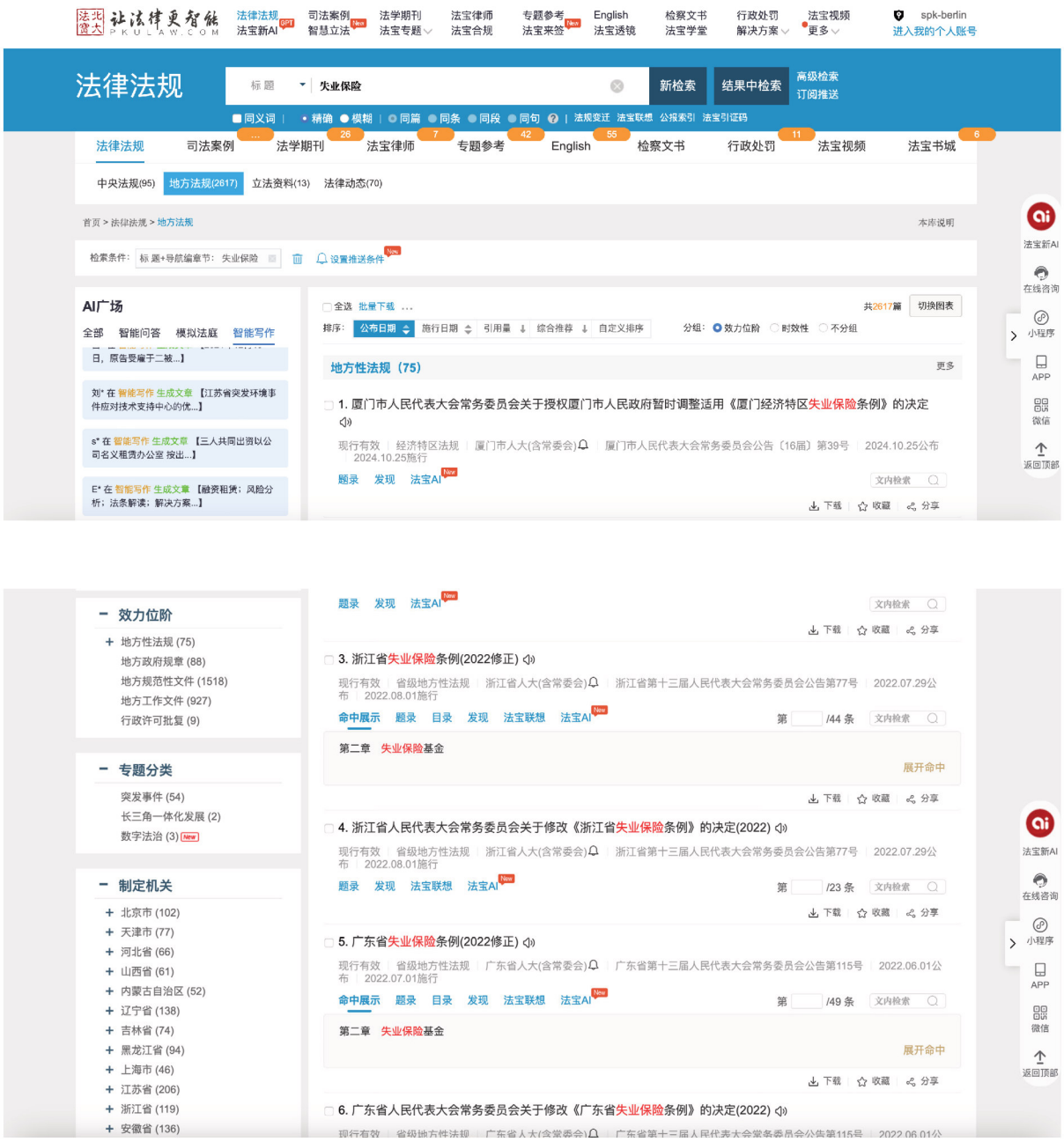
Download options include word and txt formats, and the latter was the preferred option in this project. Bulk downloads are available, but available volume thresholds differ depending on the contract with the institution providing access.

3. Setting up the database

MS-Access constitutes a comparatively simple solution for relational databases. Its core advantages are flexibility in amending existing data; integration of SQL and Graphical User Interfaces (GUIs) for queries; and integrated forms that allow convenient reading and coding of textual data. The software was able to handle the size of the datasets used here: a scope of about 2,000 documents and 50,000 sentences – as in unemployment insurance[3] – caused no problems. A scope of about 30,000 documents and 1 million sentences – as in health insurance – occasionally caused query execution to be slow. Chinese character encoding provided no significant challenges, and the flexibility of MS-Access in adding variables to existing data tables facilitates coding and cleaning the data.

3 Data was collected in early 2023, when 1,935 documents (including 89 central documents) were available. As Screenshot 1 shows, by November 2024, the number had increased to 95 central documents and 2,617 sub-national documents. The increase is to a large extent due to documents released in 2023 and 2024, as well as various 2022 documents not yet available for download in early 2023. However, there also appear to be additions in earlier years as well.

Screenshot 1. PKULAW.COM interface (2024/11/12)



3.1 The documents table

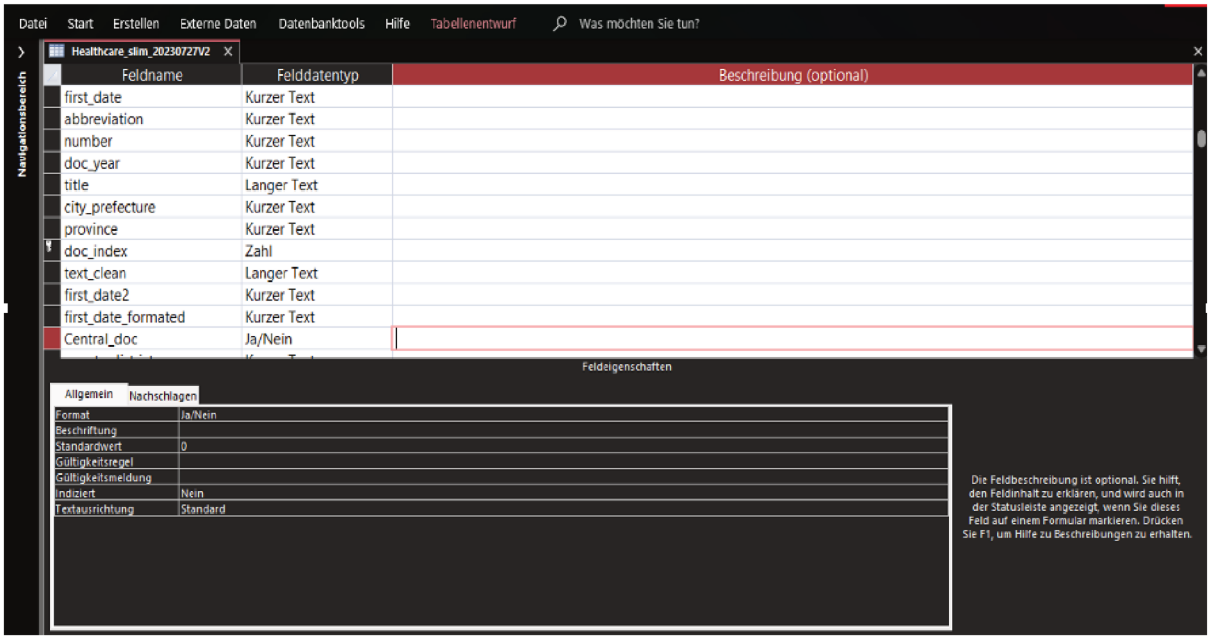
To set up the database, the text files have to be imported into a single data frame. In this project, this was accomplished with R, relying on the tidyverse and readtext packages. The code is available in file 3_1_a_txt_to_csv.R in the repository. It creates the first version of the document corpus with a column for the PKUlaw document ID and a column for the text extracted from the files.

Subsequently, core information was extracted from the text fields, using regular expressions and the tidyverse and tidytext packages in R. The code is available in the file 3_1_b_extract_clean.R in the repository. It extracts the following elements:

1) pkulaw_id: the identifying number in the PKUlaw database
2) link: the PKUlaw URL
3) first_date: first date in Chinese format that is mentioned, converted to standard Latin numbers[4]
4) document identifier[5]
   a) abbreviation: enacting body and type of document
   b) number: number of the type of document enacted by the enacting body in that year
   c) year: the year in which the document was enacted
5) title: document title
6) city_prefecture: the prefectural jurisdiction enacting the document (if applicable)
7) province: the province or provincial city enacting the document (if applicable)[6]
8) doc_index: unique identifier for each document

The information extracted needed further quality checks. For example, some documents did not feature a date, and the extracted dates and document identifiers were sometimes misleading.

Screenshot 2. Document table – design view

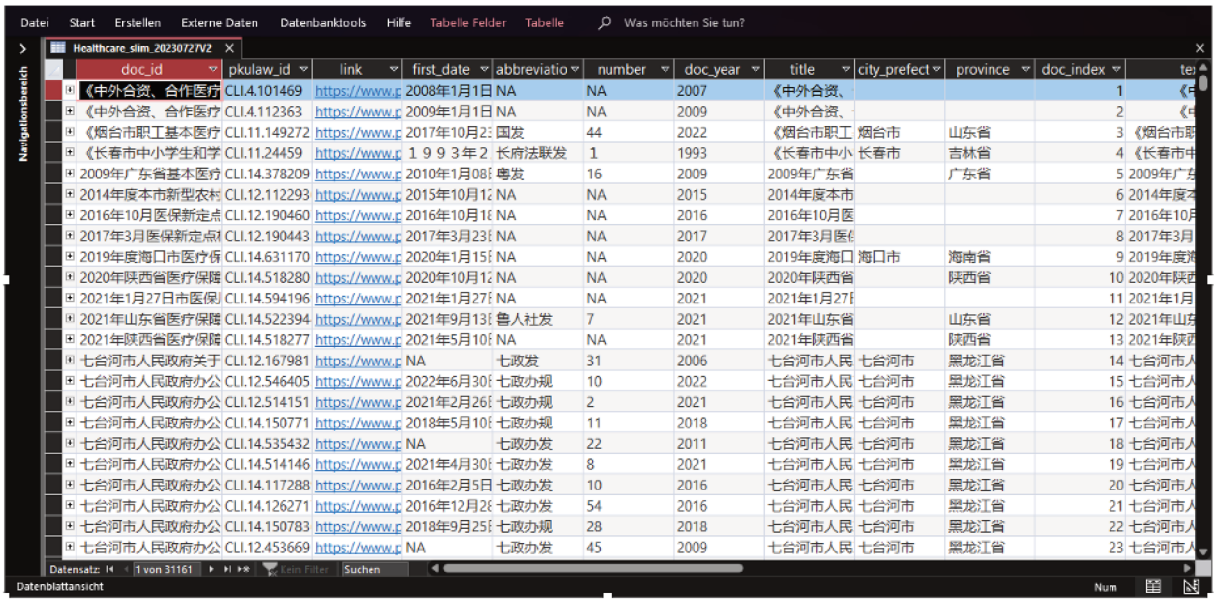


4 The first date mentioned is an indicator for the year in which the document was enacted, but it may also refer to another year.

5 Most documents come with an identifier that follows the title. It consists of an abbreviation for the enacting body and the type of document, the year of enaction, and the number of the type of document enacted in that year by that body. Example: （七政办发〔2011〕38号）. In some cases, however, documents come without an identifier, and the extracted identifier does not belong to the document processed, but to another document being cited in the processed document. Therefore, some documents may feature either no information, or misleading information. Thus, additional checks are in order.

6 Autonomous Regions were added later via a SQL query.

Screenshot 3. Document table – data view



The results were saved as a .csv file, and then the corpus was imported as a data table into MS-Access. The standard option is to import via the GUI: under external data, there is an option to import from text files. In the process, special attention must be devoted to the encoding of Chinese characters, and the correct specification of the data formats for the different columns (especially: Memo for the text and cleaned text). The Document ID (doc_index) was set as the key for the table, and there were no errors when importing the data. Screenshot 2 and 3 illustrate the imported document table in design view (variables and data types) and datasheet view respectively.

3.2 Document cleaning and preparation

Once the prepared corpus of documents has been imported into MS-Access, additional data cleaning and data preparation was required. These steps were often automated via SQL action queries, but some manual work was also required.[7] Two additional columns were created in the document table at this point: a Boolean True/False indicator, which captures whether or not the document was enacted at the central level (Central_doc); and a short text field for the name of a county-level jurisdiction (county_district), should the document have been enacted by a county-level organ.

3.2.1 Administrative jurisdictions

Data cleaning for administrative jurisdictions was relatively straightforward. The main steps were to standardize jurisdiction names, and to check the document titles regarding details of administrative jurisdiction.

First, a general select query can provide an overview over the core columns that require cleaning. It enables visual inspection, the spotting of errors, and direct correction if necessary. The filtering options in the GUI allow convenient detection, for example, of different ways of writing the name of a provincial jurisdiction (like 广西自治区 and 广西壮族自治区).

7 It is advisable to keep a back-up copy of the database before running queries that change the data.

Select Query 1. Overview and cleanup

```
SELECT Healthcare_slim_20230727V2.doc_index, Healthcare_slim_20230727V2.abbreviation,
    Healthcare_slim_20230727V2.doc_year, Healthcare_slim_20230727V2.number,
    Healthcare_slim_20230727V2.Central_doc, Healthcare_slim_20230727V2.province,
    Healthcare_slim_20230727V2.city_prefecture, Healthcare_slim_20230727V2.county_district,
    Healthcare_slim_20230727V2.title, Healthcare_slim_20230727V2.text_clean,
    Healthcare_slim_20230727V2.first_date_formated
FROM Healthcare_slim_20230727V2;
```

Some checks are suitable for automation via SQL action queries. For example, processing data in R leaves the value NA in fields with missing values, but this value does not have the same function in Access. NAs can be removed with an action query in the following form:

Action Query 1. Remove NAs from province column

```
UPDATE Healthcare_slim_20230727V2 SET Healthcare_slim_20230727V2.province = NULL
WHERE Healthcare_slim_20230727V2.province='NA';
```

Furthermore, the Autonomous Regions have not been added during data preparation with R. If the SQL query below detects the word "自治区" in the headline, it adds the name of the Autonomous Region to the province column, provided that column was previously empty. A new column for county-level jurisdictions can be filled with the names of counties in a similar way, but also requires some manual work – for example for urban county-level districts.

Action Query 2. Add Autonomous Regions to province column

```
UPDATE Healthcare_slim_20230727V2 SET Healthcare_slim_20230727V2.province =
    Left(Healthcare_slim_20230727V2.title,InStr(1,Healthcare_slim_20230727V2.title,'自治区')+2)
WHERE InStr(1,Healthcare_slim_20230727V2.title,'自治区')>0 And (Healthcare_slim_20230727V2.province Is Null Or
    Healthcare_slim_20230727V2.province='');
```

Also, the new column for central documents can be largely filled automatically. The SQL query below sets the respective field to True if the title starts with the name of any of the listed central entities.

Action Query 3. Fill central document column

```
UPDATE Healthcare_slim_20230727V2 SET Healthcare_slim_20230727V2.Central_doc = True
WHERE Healthcare_slim_20230727V2.Central_doc=False And (
    Healthcare_slim_20230727V2.title Like '人力资源和社会保障部*'
    Or Healthcare_slim_20230727V2.title Like '人力资源社会保障部*'
    Or Healthcare_slim_20230727V2.title Like '劳动和社会保障部*'
    Or Healthcare_slim_20230727V2.title Like '工业和信息化部*'
    Or Healthcare_slim_20230727V2.title Like '退役军人部*'
    Or Healthcare_slim_20230727V2.title Like '食品药品监管总局*'
    Or Healthcare_slim_20230727V2.title Like '教育部*'
    Or Healthcare_slim_20230727V2.title Like '卫生部*'
    Or Healthcare_slim_20230727V2.title Like '中央*'
    Or Healthcare_slim_20230727V2.title Like '医保局*'
    Or Healthcare_slim_20230727V2.title Like '民政部*'
    Or Healthcare_slim_20230727V2.title Like '审计署*'
    Or Healthcare_slim_20230727V2.title Like '财政部*'
    Or Healthcare_slim_20230727V2.title Like '劳动部*'
    Or Healthcare_slim_20230727V2.title Like '中共中央*'
    Or Healthcare_slim_20230727V2.title Like '司法部*'
    Or Healthcare_slim_20230727V2.title Like '全国*'
    Or Healthcare_slim_20230727V2.title Like '国家医保局*'
    Or Healthcare_slim_20230727V2.title Like '中国保监会*'
```

```
Or Healthcare_slim_20230727V2.title Like '国家医疗保障局*'
Or Healthcare_slim_20230727V2.title Like '劳动保障部*'
Or Healthcare_slim_20230727V2.title Like '卫生健康委*'
Or Healthcare_slim_20230727V2.title Like '国务院*'
Or Healthcare_slim_20230727V2.title Like '国家卫生健康委*'
Or Healthcare_slim_20230727V2.title Like '国家卫生计生委*'
Or Healthcare_slim_20230727V2.title Like '国家中医药管理局*'
Or Healthcare_slim_20230727V2.title Like '中国保险监督管*'
Or Healthcare_slim_20230727V2.title Like '国家药监局*'
Or Healthcare_slim_20230727V2.title Like '中国银保监会*'
Or Healthcare_slim_20230727V2.title Like '国家税务总局*');
```

3.2.2 Year of enactment

Cleaning the information on the year in which the document was enacted is a somewhat more complex and time-intensive process. Date verification frequently requires reading the actual text of the document, and comparing it to the first date mentioned and the date extracted from the document identifier. The extracted document identifier in most cases identifies the respective document. But occasionally, a document lacks an identifier of its own, while it also cites another document with its respective identifier, which then is falsely attributed to the document in question. For example, a local directive enacted in 2005 may lack an identifier of its own, but cite a national document enacting health insurance in 1998, the identifier of which was then attributed to the 2005 document. Furthermore, it may have set a deadline for policy implementation in 2007, which was identified as the first date mentioned. So, while the document was enacted in 2005, the year extracted from the identifier was 1998, and the first date mentioned was in 2007. There is thus no full certainty that the years extracted are the same and correspond to the year of enactment of the document.

The timing and extent of date cleaning depends on the data needs of the researcher. In any case, adding a field "date verified" to the document table can help keeping track of the verification progress across different projects. In the workflow discussed in the following section, the dates were checked while reading the documents and reconstructing the implementation process. In addition, the year extracted from the identifier is either smaller than or equal to the year of enactment, because documents may cite other documents from the past, but not from the future. For a comprehensive cleaning operation, the following approaches can help identify documents with inaccurate date information via SQL queries:

1) The year extracted from the identifier is not the same as the first year mentioned
2) The document was issued by a sub-national organ, but the abbreviation is that of a central organ
3) The document mentions multiple years in Latin or Chinese characters

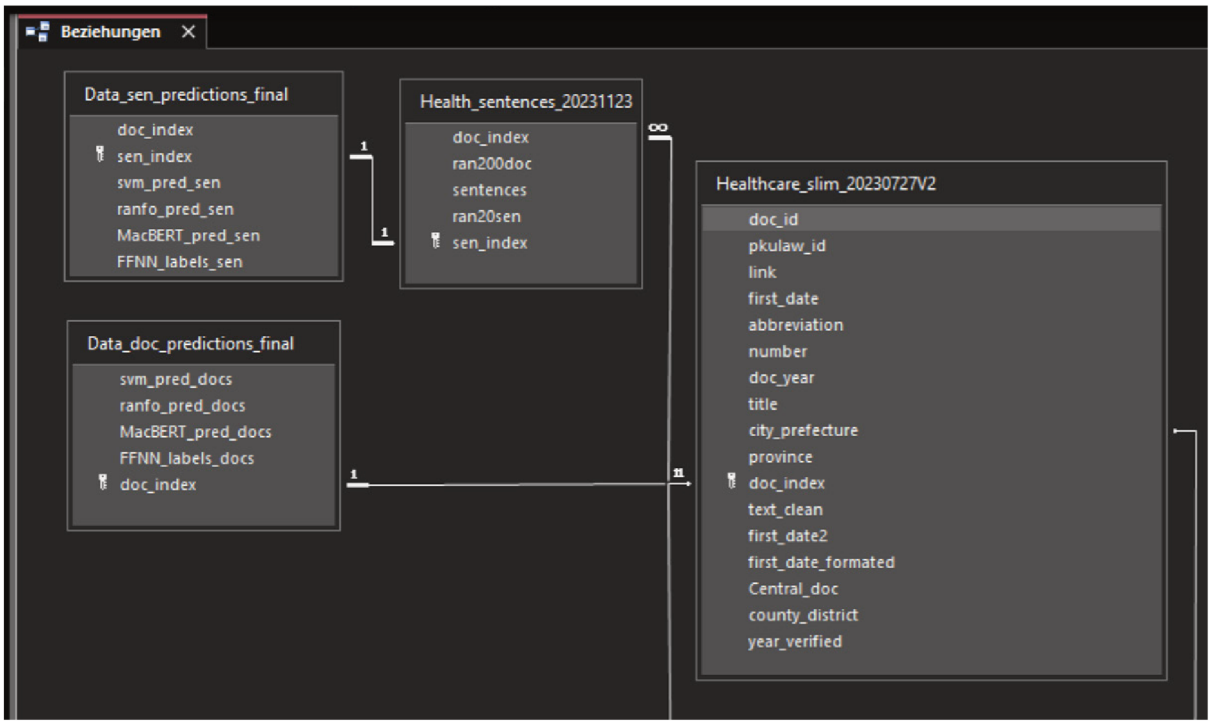
These approaches can be combined with keyword search where applicable to reduce the workload in a given project.

3.3 The sentences table and data to be labelled

The documents were furthermore disaggregated into sentences, which offers two advantages: First, identifying relevant sentences via queries facilitates a quick overview over the regulatory situation on a given topic. Second, text classification with machine learning can be conducted at sentence level, which allows for a large number of observations to be coded in a reasonable amount of time. Each sentence was assigned a unique identifier (sen_index), and a separate sentence table was integrated into the database.

Machine learning requires a representative subset of the sentences to be extracted and labelled. Extracting a representative subset requires randomization across documents and sentences. At document level, a random variable with 200 different values was created before the split into sentences. At

Screenshot 4. Relationship manager



sentence level, a random variable with 20 different values was created after the split.[8] Subsequently, 21 batches of training data were created by iterating over both random numbers 3 steps at a time. This way, each batch included about 1,000 sentences. The code for the implementation of the split and the selection of training data is provided in 3_3_split_to_sentences.R in the repository.

The sentences were subsequently loaded into the database with the standard data import routine, as described for the documents above. With the document identifier (doc_index), the sentences were linked to the documents in the relationship manager in the database tools section. A one-to-many relationship between documents and sentences was created with referential integrity. This step is important, among other things, to aggregate results from sentence-level text classification at document level, or to create a query which merely displays sentences considered relevant in a given document. The sentence identifier (sen_index) later allows to connect the sentences to the respective predictions.

4. Text mining with SQL

This section discusses how SQL queries help create variables from administrative documents, which can be used for quantitative analysis in combination with survey data or official time series by the National Bureau of Statistics. An important pre-condition is a match in the level of analysis between the textual data and the statistical data. As noted above, PKUlaw provides a fairly comprehensive collection of documents at provincial and prefectural level, while coverage of county-level jurisdictions is more sporadic.

The example in this section focuses on a complementary health insurance program: Catastrophic Medical Insurance (CMI), which complements the Urban and Rural Residents' Basic Medical Insurance. It is usually managed and pooled at the level of prefectural or provincial city. Corresponding hospitalization data was extracted from the China Health and Retirement Longitudinal Study (CHARLS),

8 There are 2 principal options: randomization across all sentences (applied for the unemployment insurance data), and randomization across the sentences within each document (applied for the healthcare data).

which published its sample of prefecture-level jurisdictions. The CMI program was enacted and extended in China between 2012 and 2016.

The aim was to derive a binary variable about the implementation status in each prefecture and year from the documents. Select Query 2 identifies relevant documents by checking for the presence of relevant key words in the title or the text of the documents. The list of key words was amended several times, because new variations in the name of the policy or relevant equivalents came up during the analysis. The query listed here automatically excludes documents enacted after 2016, and Screenshot 5 provides an overview of the results.[9]

Select Query 2. Catastrophic Medical Insurance documents, 2016 or earlier

```
SELECT
    Healthcare_slim_20230727V2.doc_index,
    Healthcare_slim_20230727V2.province,
    Healthcare_slim_20230727V2.city_prefecture,
    Healthcare_slim_20230727V2.county_district,
    Healthcare_slim_20230727V2.doc_year,
    Healthcare_slim_20230727V2.number,
    Healthcare_slim_20230727V2.title
FROM
    Healthcare_slim_20230727V2
WHERE
    If(IsNumeric(Healthcare_slim_20230727V2.doc_year),
        CInt(Healthcare_slim_20230727V2.doc_year), 0) <= 2016
    AND (
        Healthcare_slim_20230727V2.title Like '*大病保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大病保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*补充医疗保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大病商业保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大病补充保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大病医疗保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大额医疗保险*'
        OR Healthcare_slim_20230727V2.text_clean Like '*大病补充补偿*'
);
```

The reconstruction of the implementation process proceeded in three steps. First, to read the central documents pertaining to policy implementation and reconstruct the process at central level. Second, to read the documents at provincial level and reconstruct the process in the provinces. Third, to read the documents for those prefectural jurisdictions that were also sample jurisdictions for the CHARLS survey, which provided the statistical data for this study. The second and third step were mostly combined by filtering for one province at a time in the GUI of Select Query 2. The documents could be read in a corresponding form (see: Screenshot 6), and missing or wrong year information could be adjusted when needed. The implementation dates were collected in an Excel file, with one sheet for provinces and prefectural jurisdictions respectively, and with years in the columns and jurisdictions in the lines. For every jurisdiction, the ID and year of enactment of the document enacting implementation were listed in the year of implementation.[10] When the analysis was finished, an additional sheet for the sample jurisdictions was filled with the numerical data, and subsequently added to the data set for quantitative analysis.

9 For the actual study, the year selection was conducted via the GUI; each province was analyzed separately, and later years were added for additional context if the previous documents provided ambiguous results.

10 Occasionally, additional online search or consultation of documents enacted after 2016 was necessary to clear up ambiguities. Furthermore, some additional documents were found and added to the database via the reading form, increasing the total number of documents from the original 31,139 to 31,161.

Screenshot 5. Select Query 2 output

QRYcmi_01a_DBBX_implementation

doc_index	province	city_prefect	county_dist	doc_year	number	title
14	黑龙江省	七台河市		2006	31	七台河市人民政府关于印发《七台河市城镇职工基本医疗保险办法》的通知
20	黑龙江省	七台河市		2016	10	七台河市人民政府办公室关于印发七台河市开展新农合大病商业保险工作实施方案的通
21	黑龙江省	七台河市		2016	54	七台河市人民政府办公室关于印发七台河市整合城乡居民基本医疗保险工作实施方案的
26	黑龙江省	七台河市		2016	45	七台河市人民政府办公室关于贯彻落实《黑龙江省城乡医疗救助暂行办法》的实施意见
27	海南省	万宁市		2014	108	万宁市人民政府办公室关于印发万宁市2015年度新型农村合作医疗参合金征收工作方案
28	海南省	万宁市		2015	127	万宁市人民政府办公室关于印发万宁市2016年度新型农村合作医疗参合金征收工作方案
29	海南省	万宁市		2016	121	万宁市人民政府办公室关于印发万宁市2017年度新型农村合作医疗参合金征收工作方案
31	海南省	万宁市		2016	115	万宁市人民政府办公室关于印发万宁市建档立卡贫困人口医疗保障扶贫工作实施方案的
40	海南省	三亚市		2013	101	三亚市人民政府关于印发三亚市城乡重特大疾病医疗救助实施方案的通知
64	福建省	三明市		2013	82	三明市人民政府关于印发2013年新型农村合作医疗市级统筹管理实施方案的通知
67	福建省	三明市		2010	1	三明市人民政府办公室关于做好2010年新型农村合作医疗工作的通知
68	福建省	三明市		2008	57	三明市人民政府办公室关于印发三明市区农民工大病医疗保险实施意见的通知
71	福建省	三明市		2008	151	三明市人民政府办公室关于印发三明市城镇个体工商户业主及其雇工参加城镇职工基本
72	福建省	三明市		2010	84	三明市人民政府办公室关于印发三明市城镇职工基本医疗保险市级统筹实施办法的通知
73	福建省	三明市		2009	118	三明市人民政府办公室关于印发三明市大学生参加城镇居民基本医疗保险实施办法的通
83	福建省	三明市		2013	7	三明市人民政府办公室转发市发展改革委等单位关于进一步完善城乡居民大病保险工作
85	福建省	三明市		2011	74	三明市人民政府批转市民政局等部门关于三明市城乡医疗救助实施办法的通知
91	河南省	三门峡市		2008	30	三门峡市人民政府关于印发三门峡市城镇居民基本医疗保险暂行办法和大额医疗费补充
92	河南省	三门峡市		2010	3	三门峡市人民政府关于印发三门峡市破产企业退休人员医疗保险暂行办法的通知
93	河南省	三门峡市		2008	30	三门峡市城镇居民基本医疗保险暂行办法
95	上海	上海市		2014	15	上海保监局关于上海市保险公司第三批大病保险经营资质名单的公告
96	上海	上海市		2014	6	上海保监局关于上海市城乡居民大病保险业务经营资质管理有关问题的通知
98	上海	上海市		2014	171	上海保监局关于印发《上海市保险公司城乡居民大病保险投标管理暂行办法》的通知
101	上海	上海市		2015	12	上海市人力资源和社会保障局、上海市医疗保险办公室关于本市基本医疗保险2015医保
111	上海	上海市		2016	16	上海市人力资源和社会保障局、上海市医疗保险办公室关于本市基本医疗保险2016医保
118	上海	上海市		2016	42	上海市人力资源和社会保障局、上海市医疗保险办公室、上海市教育委员会等关于做好
172	上海	上海市		2012	22	上海市人力资源和社会保障局、上海市医疗保险办公室关于本市基本医疗保险2012医保
182	上海	上海市		2012	390	上海市人力资源和社会保障局、上海市医疗保险办公室关于小城镇基本医疗保险门急诊
189	上海	上海市		2013	18	上海市人力资源和社会保障局、上海市医疗保险办公室关于本市基本医疗保险2013医保
196	上海	上海市		2014	9	上海市人力资源和社会保障局、上海市医疗保险办公室关于本市基本医疗保险2014医保
198	上海	上海市		2014	37	上海市人力资源和社会保障局、上海市医疗保险办公室、上海市教育委员会等关于做好
214	上海	上海市		2016	496	上海市人力资源和社会保障局、上海市医疗保险办公室、上海市卫生和计划生育委员会

Datensatz: 1 von 2940 | Kein Filter | Suchen

Screenshot 6. Reading form

FRM_doc_read

doc_index: 15209 province: 四川省 city_prefecture: 成都市 doc_year: abbreviation: NA

title: 成都市医疗保障局关于《成都市第十七届人大二次会议第592号建议》答复的函 number: NA

text_clean: Central_doc year_verified first_date: NA

成都市医疗保障局关于《成都市第十七届人大二次会议第592号建议》答复的函成都市医疗保障局关于《成都市第十七届人大二次会议第592号建议》答复的函刘益民委员： 您提出的《关于适度增补特病门诊种类科学配置医保资金使用的建议》（第592号）已收悉，现答复如下： 一、关于我市医保门诊医疗费用报销政策 我市城镇职工基本医疗保险制度自2001年起施行，在设立之初就确定基本医保险基金由个人账户和统筹基金构成，统筹基金用于支付住院医疗费，个人账户用于支付门诊医疗费。但因个人账户金额有限，为保障参保人员医疗待遇，减轻门诊个人负担，我市把部分需长期门诊治疗的慢性疾病、重特大疾病纳入了门诊特殊疾病管理，由统筹基金支付。目前纳入门诊特殊疾病范围的有四类疾病33个病种，比其他统筹地区种类更多，报销政策更优越。除门诊特殊疾病外，我市还有其他一门诊医疗费用可纳入统筹基金支付范围：城镇职工有家庭病床、门诊抢救无效死亡发生的医疗费用、入院前3日内的阳性特殊检查等；城乡居民有门诊统筹和一般诊疗费等。 二、关于约定符合医保报销的门诊类疾病的金额 目前，为控制医疗费用过快增长，我市医在对医疗机构支付门诊特殊疾病医疗费用时，采取部分病种限额和定额付费。医疗机构应在疾病治疗时主动控制不合理费用的产生，超限额、定额标准的费用由医疗机构承担。限额、定额付费未针对参保患者，原因是慢性疾病的转归特点是并发症、合并症复杂，治疗药种类繁多、新型诊疗技术更新较快，无法科学合理的对每个病种建立对应的治疗药品、诊疗项目医保可报销目录，从而无法确定合理的销金额。此外，我市参保人员可享受基本医疗保险和大病医疗互助补充保险两种待遇，城镇职工基本医疗保险报销比例最低可达85%，大病医疗互助补充保险按级距报销，最低段的报销比例也达到了77%，如果医保按病种确定报销金额，需要重新评估医保基金（资金）承受能力和各病种的平均治疗水平，情况较复杂，且可能造成部分参保人员门诊特殊疾病医保待遇降低。 三、关于门诊与住院的相互转和对医疗机构的监督和引导 医疗保险基金由参保单位和个人缴纳，并非国家投入，基本原则是以收定支、收支平衡、略有结余。医部门在制定政策时，应充分考虑医保基金的可承受能力和使用的可持续性，必须坚持基本保障、稳健持续、依法依规切实维护人民群众本医疗保障需求，同时也要实事求是、量力而行。一方面作为医疗保障供给方的医疗机构，多数以获取最大利益为目的，所以入院指征低、医疗价格虚高，医药乱象，大处方、滥检滥查等现象突出，导致住院人次和医疗费用居高不下，且逐年上涨；另一方面，作为需求的参保患者，因医疗知识欠缺，就医需求盲目，容易被医疗机构导向影响，无法自主控制不合理医疗费用的产生。而作为医疗服务购买的医保部门，既要保障参保人员合理合法的医疗需求、减轻个人经济负担，又要在维护基金安全、杜绝医疗乱象等方面不断努力。从我实际情况来看，门诊特殊疾病政策出台，减轻了个人门诊医疗负担，但在降低住院率的效果上并不明显，这和前面所述两方面问题不无系。正如您所了解的，我市医保部门已经采取了很多综合措施来规范医疗服务和管理医保基金，如总额控制、病种限额、智能审核、专稽核、大数据监控分析等。如何在提高人民群众获得感的同时，保障基金安全和合理支出，扎实推进医疗保障制度管理规范化、标准化法治化，是全国甚至世界医疗保障面临的难题，需要由政府主导，医保、卫生和其他相关部门的相互协作、共同承担。 目前，国家保局正在准备建立全国统一的医疗保障待遇清单，将在全国范围内对住院、普通门诊、门诊慢特病的待遇支付进行具体规定，科学确定保待遇水平、完善风险分担机制，鼓励发展多层次医疗保障体系。下一步，我市将根据国家局的指导，完善医保相关政策。 感谢您市医保局工作的关心和支持，希望继续提出宝贵意见和建议，您对以上答复有什么意见，请填写在《办理情况反馈意见表》上，以便我及时改进。 （联系人：梁树；联系电话：87706731）

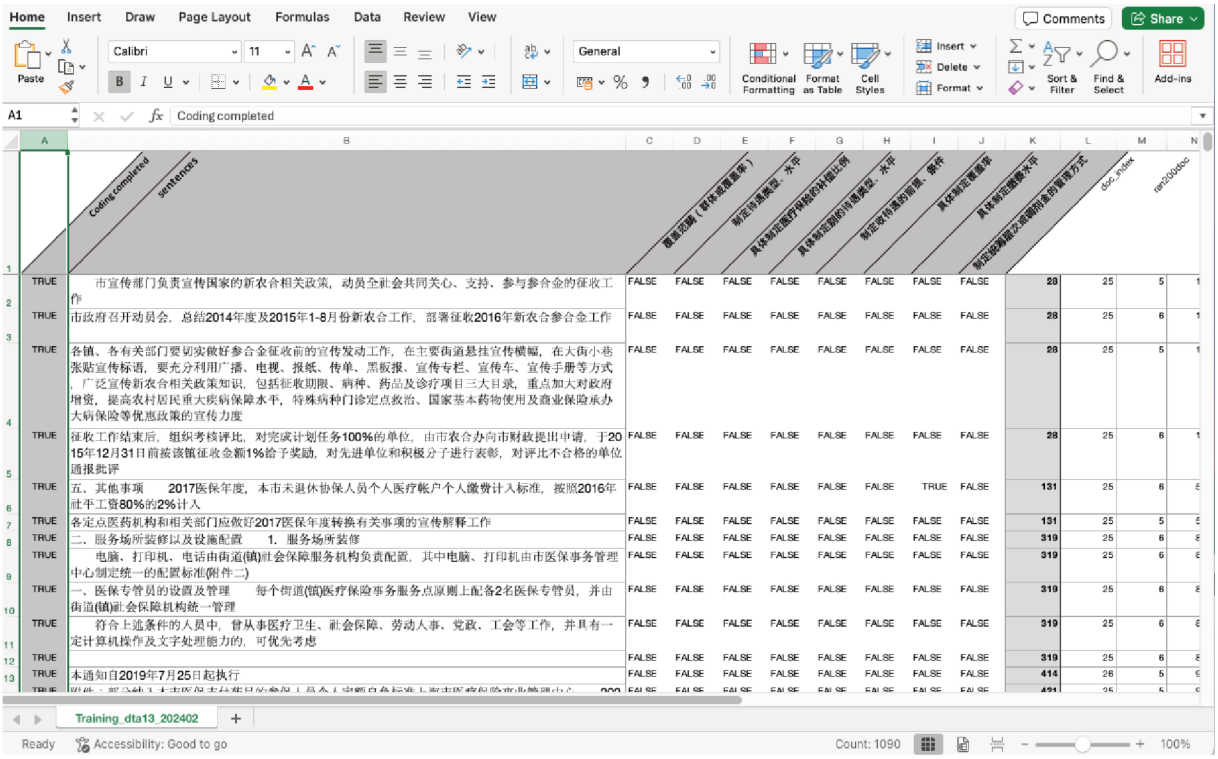
5. Document classification with machine learning

This section provides an overview of the text classification workflow, which is applied in ongoing research on pension, health and unemployment insurance in China. The complementary code files are in Python, but machine learning algorithms in R should allow for largely comparable results.

5.1 Training data labelling and control

The batches of training data described above were labelled by two student assistants. The sentences were checked regarding whether or not they contained information pertaining to the inclusiveness or the scope of benefits of health insurance. Excel files provide a convenient format for labelling. The files contained a column for each variable to be extracted, and a field to check once a sentence has been labelled. Fields not intended for editing (like the sentence text or the sentence ID) were formatted as locked and protected with a password. Subsequently, the labelled batches were integrated in a single file.

Screenshot 7. Training data labelling



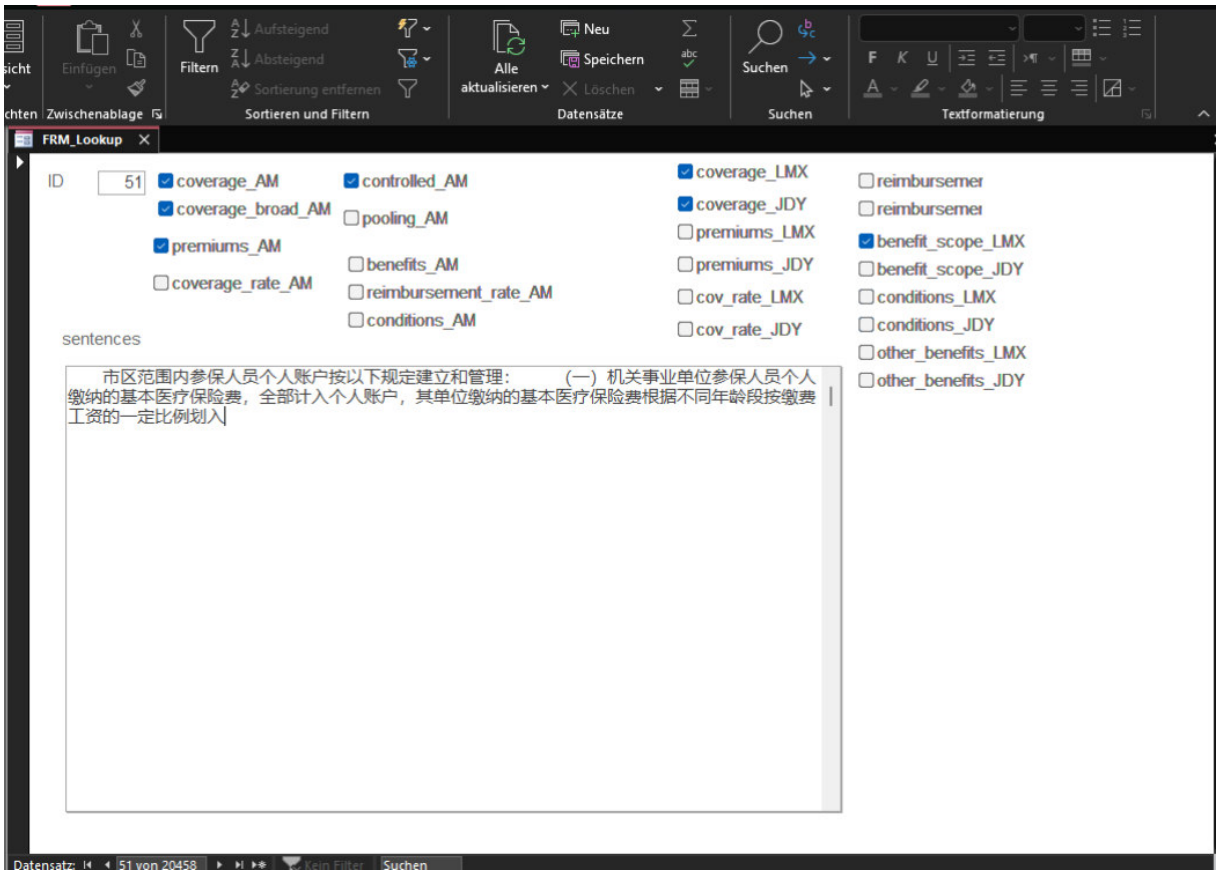
Select Query 3. Labelled sentences that have not been controlled yet

```
SELECT Train_dta_240910.ID, Train_dta_240910.controlled_AM,
    Train_dta_240910.coverage_LMX, Train_dta_240910.coverage_AM,
    Train_dta_240910.coverage_broad_AM, Train_dta_240910.coverage_JDY,
    Train_dta_240910.cov_rate_LMX, Train_dta_240910.cov_rate_JDY,
    Train_dta_240910.premiums_LMX, Train_dta_240910.premiums_JDY,
    Train_dta_240910.sentences, Train_dta_240910.pooling_JDY,
    Train_dta_240910.pooling_LMX, Train_dta_240910.benefit_scope_JDY,
    Train_dta_240910.benefit_scope_LMX, Train_dta_240910.other_benefits_JDY,
    Train_dta_240910.other_benefits_LMX, Train_dta_240910.conditions_JDY,
    Train_dta_240910.conditions_LMX, Train_dta_240910.reimbursement_rate_AM,
    Train_dta_240910.benefits_AM, Train_dta_240910.coverage_rate_AM,
```

```
    Train_dta_240910.conditions_AM, Train_dta_240910.premiums_AM,
    Train_dta_240910.pooling_AM
FROM Train_dta_240910
WHERE (((Train_dta_240910.controlled_AM)=False)
    AND ((Train_dta_240910.coverage_LMX)=True))
    OR (((Train_dta_240910.controlled_AM)=False)
    AND ((Train_dta_240910.coverage_JDY)=True));
```

The labelled data was loaded into a small, separate Access database to check the labels. Access allows a convenient combination of queries and forms, which facilitates the control of relevant observations. Select Query 3 selects all the relevant columns for labelling the different variables, for observations that were labelled as containing information about coverage rules by at least one of the two student assistants. Screenshot 8 illustrates a form, which displays all the relevant fields alongside the labelled sentences and allows for quick detection and correction of false positives (for example, a sentence that does not include information about coverage rules, but which was coded as including coverage rules by at least one of the student assistants. All positive labels were checked. The form was adjusted to only show data listed in Select Query 3. Once false positives were corrected, targeted queries looking for common keywords of the true positives can help identify sentences with false negatives. Here, too, a special form for working through this query can be set up.

Screenshot 8. Encoding form



In the process, two coverage indicators were created: a narrow one and a broad one. The narrow version (y_narrow) focuses on statements directly setting rules of inclusion. The broader version (y_broad) also includes statements that indirectly point to the inclusion of a group – for example by setting specific premium levels for migrant workers, implying their coverage by the system. Both indicators displayed an

imbalanced distribution: out of 20,264 coded sentences,[11] only 1,150 discussed coverage in the broad sense, and only 418 discussed coverage in the narrow sense. Such strong imbalance is a challenge to the performance of machine learning models. Machine learning focused on the broad indicator, and the more complex type of information it transmits. To mitigate the imbalance, the training data was up-sampled to achieve an equal balance between positive and negative observations, whereas the test data retained the imbalanced distribution. Once the controls were finished, data was exported via a query that selected the required columns: the sentences, the labels, and the identifier for the documents (doc_index) and the sentences (sen_index). The data was saved as a .csv file.

5.2 Preparing the predictors

For conventional machine learning models and Feed-Forward Neural Networks, the Bag-of-Words approach was applied. Bag-of-Words means that words – or tokens – are removed from the context of word order and grammar, and treated independently. The results of the analyses improve when noise is removed and the focus remains on meaningful words and tokens. The sentences were thus prepared in several steps outlined below, before the models were trained. By contrast, Transformers such as BERT can directly process full sentences, to which they apply their own encoding routines.[12]

For the Bag-of-Words approach, first, the data needs to be segmented and cleaned. Segmentation is crucial, because the Chinese language does not separate words with spaces. It creates tokens by separating the text into word units, retaining spaces, punctuations and numbers as tokens. Subsequently, word-units that carry little meaning were removed, such as the ideographic space ('\u3000'), punctuation, as well as stop words and numbers. These steps were applied to both, the labelled data and the unlabelled data. Subsequently, the labelled data was split into training and test data. These steps are found in the file 5_2_a_clean_split.ipynb. The training data contained 13,576 sentences, and the test data contained 6,688 sentences.

Second, vectorization encodes the textual data, creating numerical data suitable for machine learning. Three approaches to vectorization were applied here: first, Multihot Encoding creates a binary variable indicating the presence of each word in the vocabulary in a given sentence. Multihot encoded data had 18,559 features. Second, the Term-Frequency/Inverse-Document-Frequency (TF-IDF) indicator creates a numerical variable that considers the frequency of a term in a sentence and in the corpus as a whole. High values are assigned to words that feature often in a sentence, but are rare in the corpus. The TF-IDF vectorizer was set to extract unigrams and bi-grams appearing in at least 3 documents and less than 80% of documents of the training data. It yielded 21,960 features. Finally, Meta's fastText embeddings are a form of word embeddings that convert the entire sentence into 300 numerical features, which capture the meaning of the given sentence in a standardized way.[13] The code is provided in the files 5_2_b_vectorization.ipynb and 5_2_c_Fasttext_vectorization.ipynb.

5.3 Pre-processing, training and prediction

Models were trained in four larger classes of algorithms: Support Vector Machines, Tree-based models, Feed-Forward Neural Networks, and Transformers. These models allow for and respond differently to different vectorization approaches and different pre-processing steps, namely feature selection and dimensionality reduction. Table 1 illustrates the main characteristics of the four workflows presented

11 Originally, there were 20,458 coded sentence observations. Among them, 170 were empty, and another 24 contained only stopwords, leaving 20,264 sentences for model training.

12 For the Transformers, the necessary preparation steps are included in the respective code files.

13 Word vectors for 157 languages are available here: https://fasttext.cc/docs/en/crawl-vectors.html

here. The algorithms are presented in order of sophistication and computational requirements, with Support Vector Machines being the simplest and large Transformers the most demanding approaches. Training data was up-sampled, so that all models trained with an equal number of positive and negative observations, whereas the test data retained the original unbalanced structure.

Table 1. Vectorization and classifiers

Algorithm class	Support Vector Machine (SVM)	Tree-based	Neural Networks	
Specification	Linear, L2 penalty, C = 9	Random Forest	Feed-Forward	Transformer: Chinese MacBERT (large)
Up-sampling	Training data	Training data	Training data	Training data
Bag of Words approach	Yes	Yes	Yes	No
Vectorization	fastText	Multihot	TF-IDF	generic
Number of features	300	18,559	21,960	
Feature Selection	SVM, L1 penalty, C = 1		SVM, L1 penalty, C = 2	
Dimensionality reduction	PCA (95% variance)		PCA (95% variance)	
Remaining features	155		640	
Tuning	Grid search	Manual	Manual	
Accuracy	88%	93%	89%	94%
Sensitivity	87%	85%	89%	92%

Algorithms relying on the Bag-of-Words approach can benefit from feature selection and pre-processing of features. These preparations reduced noise in the data, and focused on the numerical inputs (fastText and TF-IDF vectorization) in combination with SVMs and Feed-Forward Neural Networks. Features with an effect on the outcome were selected via a SVM with a L1 penalty. Subsequently, the dimensionality of inputs was reduced via Principal Component Analysis (PCA), a mathematical algorithm that reduces the input matrix to a smaller number of orthogonal vectors, thereby eliminating correlated features and the potential distortion that may come with them. These steps render a drastically reduced number of features, which represent the relevant variation in the dataset. Model performance improved, but the features ceased to be interpretable.

Model evaluation focused on the metrics of Accuracy (the share of correct predictions in all observations) and Sensitivity (the share of predicted positives among all positive observations). This combination makes sense given the imbalanced nature of the coverage indicator. Accuracy alone could not guarantee sufficient numbers of correctly predicted positive cases, and false positives (observations predicted as positive despite truly being negative) are easier to control and correct than false negatives.

Predictions required higher computational capacity than model training. As noted above, the training data contained 13,576 sentences, and the test data contained 6,688 sentences. The computational requirements remained largely moderate, allowing for model training in Jupyter notebooks on a laptop with rudimentary GPU capacity for SVMs, Random Forest and Feed-Forward Neural Networks. By contrast, the complete sentence dataset included about 1 million observations,[14] and the difference in magnitude caused a substantial increase in the computational resources required. Therefore, with the exception of SVMs, actual predictions were conducted in .py files on a separate server with greater capacity. The relevant files are provided in the repository with the number-code "5_3_". The predictions

14 The original sentences table includes 1,003,136 observations. After removing missing observations and observations containing merely stop-words, 993,526 observations were left in the unlabeled data for predictions.

rendered labels of 1 and 0 at sentence level. These predictions were aggregated at document level as well, and both were subsequently imported into the Access database.

5.4 Analysis

In the Access database, getting the imported predictions ready for analysis required several steps: First, to create relationships between the sentence and document tables and the tables with the predictions via the respective unique identifiers (doc_index and sen_index; see also: Screenshot 4).

Second, to create an index from the predictions at document level (Doc_score) and to arrange the documents in descending order of the index (see: Select Query 4). The index can simply sum up the number of positive predictions of the four classifiers in each document. In smaller data sets, like for unemployment insurance, reading the relevant documents until relevant information ceases to appear is an option for qualitative analysis.[15] For larger datasets, it makes sense to create separate queries for documents issued by different administrative levels, as Select Query 4 below does, for a quicker overview and to facilitate subsequent summaries.

Select Query 4. Creating an index for prefectural-level documents and arranging documents

```
SELECT Healthcare_slim_20230727V2.doc_index, Healthcare_slim_20230727V2.province,
    Healthcare_slim_20230727V2.city_prefecture, Healthcare_slim_20230727V2.county_district,
    Healthcare_slim_20230727V2.doc_year,
    [svm_pred_docs]+[MacBERT_pred_docs]+[FFNN_labels_docs]+[ranfo_pred_docs] AS
    Doc_score, Healthcare_slim_20230727V2.title, Data_doc_predictions_final.svm_pred_docs,
    Data_doc_predictions_final.ranfo_pred_docs, Data_doc_predictions_final.MacBERT_pred_docs,
    Data_doc_predictions_final.FFNN_labels_docs
FROM Healthcare_slim_20230727V2 INNER JOIN Data_doc_predictions_final ON
    Healthcare_slim_20230727V2.doc_index = Data_doc_predictions_final.doc_index
WHERE (((Healthcare_slim_20230727V2.province) Is Not Null) AND
    ((Healthcare_slim_20230727V2.city_prefecture) Is Not Null) AND
    ((Healthcare_slim_20230727V2.county_district) Is Null))
ORDER BY ([svm_pred_docs]+[MacBERT_pred_docs]+[FFNN_labels_docs]+[ranfo_pred_docs]) DESC;
```

Third, to create a prediction index at sentence level (Sen_score), again by simply adding the predictions together. There are four predictions for every sentence, and as noted above, each classifier produced a considerable number of false positives. However, the errors of the different classifiers should cancel each other out to some extent, and the number of false positives should be lower for sentences to which multiple classifiers assign a positive label. Select Query 5 identifies sentences in which at least three of four classifiers assign a positive label.

Select Query 5. Selection query with index for sentences with at least 3 positive labels

```
SELECT Health_sentences_20231123.sen_index,
    [svm_pred_sen]+[MacBERT_pred_sen]+[FFNN_labels_sen]+[ranfo_pred_sen] AS Sen_score,
    Health_sentences_20231123.sentences, Data_sen_predictions_final.svm_pred_sen,
    Data_sen_predictions_final.ranfo_pred_sen, Data_sen_predictions_final.MacBERT_pred_sen,
    Data_sen_predictions_final.FFNN_labels_sen, Health_sentences_20231123.doc_index
FROM Health_sentences_20231123 INNER JOIN Data_sen_predictions_final ON
    Health_sentences_20231123.sen_index = Data_sen_predictions_final.sen_index
WHERE ((([svm_pred_sen]+[MacBERT_pred_sen]+[FFNN_labels_sen]+[ranfo_pred_sen])>=3));
```

Further steps of analysis can be determined depending on the specific goals and research interest. One option is to further filter the results to extract coverage regulations pertaining to a specific insurance program. Another option is to trace the policy process in specific provinces or cities, similar to the process

15 Depending on the task, the relevant documents may also be exported from Access and analyzed with a tool that offers more advanced options for qualitative coding, such as MaxQDA.

described in section 4. Yet another option is to combine these results with the search for group-specific or other keywords. Last but not least, if the predictions are sufficiently interpretable, they can also be used for quantitative analysis.

Screenshot 9. Selection query with index for sentences with at least 3 positive labels

QRYcov_pred_sen_3positive

doc_index	Sen_score	sentences
3	3	第二条　职工基本医疗保险实行市级统筹，并坚持以下原则：　　（一）医疗保险水平与本市经济发展水平和各方面承受能力相适应；
3	4	第三条　本市行政区域内的机关、企业、事业单位、社会团体、民办非企业单位及其职工（含建国前老工人），城镇个体经济组织业主及其
3	4	用人单位应按本单位在职职工工资总额的7%缴纳基本医疗保险费，在职职工按本人工资总额的2%缴纳
3	4	灵活就业人员只需缴纳划入基本医疗保险统筹基金部分的医疗保险费，不建立个人账户
3	4	第七条　破产企业应按照《中华人民共和国企业破产法》及有关规定，优先偿付欠缴的基本医疗保险费，并按烟台市上年度离休人员（含建
3	4	第八条　参加职工基本医疗保险的人员，其基本医疗保险缴费年限（含视同缴费年限）男不满25年、女不满20年的退休后不享受基本医疗保
3	3	中断缴费人员缴费中断期间不享受基本医疗保险待遇
3	3	退休人员在上述支付比例的基础上再提高5%
3	4	第十七条　个人账户金按照"效率优先、兼顾公平、适当照顾老年人"原则，暂按以下标准划入：　　35周岁以下（不含35周岁）在职职工月
3	4	在职职工和退休人员按年度一次性缴纳大额救助金，标准为每人每年36元
3	4	第三十一条　鼓励有条件的用人单位，在参加基本医疗保险的基础上，建立补充医疗保险，补充医疗保险费在工资总额4%以内的部分，可直
3	3	补充医疗保险办法，经用人单位职工代表大会或职工大会讨论通过后实施
3	4	第三十三条　离休人员、一至六级革命伤残军人的医疗保险办法按原规定执行
3	4	第三条　本市行政区域内不属于职工基本医疗保险参保范围的城乡居民，均应依法参加居民基本医疗保险
3	4	教育部门负责组织在校学生（幼儿园）统一参加居民基本医疗保险，做好参保登记、代收代缴、费用结算等工作
3	4	民政、残联部门负责做好孤儿、特困人员、城乡最低生活保障对象、享受定期定量救济的60年代精减退职老职工、重度残疾人等特殊群体（
3	3	参保居民按规定缴纳基本医疗保险费，享受相应的医疗保险待遇
3	3	（一）个人缴费标准分两档，2018年缴费标准为：一档每人每年230元，二档每人每年380元
3	4	（三）各类在校学生和其他未成年居民（以下统称未成年居民）缴费标准为：各类在校学生个人缴费标准暂执行2017年标准，其他未成年居
3	4	各级政府（管委）应按相关规定对特殊群体和建档立卡贫困人口个人缴费部分给予资助
3	3	第八条　居民基本医疗保险实行年缴费制度，每年9月1日至12月31日为下一年度参保缴费期，参保居民应于参保缴费期内缴纳下一年度基本
3	4	（一）在校学生按学籍以学校为单位组织参保登记和缴费，其他居民以户为单位由其户籍所在地或居住地乡镇政府（街道办事处）组织参保
3	4	（二）特殊群体、建档立卡贫困人口由民政、残联部门和扶贫办统一组织参保登记和缴费
3	4	（三）新生儿自出生之日起90日内办理参保手续并缴纳出生当年居民基本医疗保险费，自出生之日起享受居民医疗保险待遇
3	3	（三）未成年居民享受二档缴费的医疗保险待遇
4	4	《长春市中小学生和学龄前儿童住院医疗保险办法》实施细则《长春市中小学生和学龄前儿童住院医疗保险办法》实施细则（１９９３年２月９
4	4	第二条　《办法》第二条规定的被保险人系指：在我市行政区域内的所有中、小学校（含职业学校、特殊学校）的在册学生和一周岁以上学
4	4	被保险人参加本保险后因病、伤休学后，其所在学校保留学籍的学生，仍可作为本保险的被保险人，但年龄一般不超过十八周岁
4	4	盲聋哑学校及弱智学校学生年龄应不超过二十周岁
4	4	第三条　本保险以被保险人所在学校或幼儿园（街道办、村委会，下同）为投保代办单位（以下简称代办单位），统一于每学年第一学期开
4	3	代办单位是学校、幼儿园的必须确保９０％以上的投保率；代办单位是街道办的，必须确保８５％以上的投保率；对未达到投保率的，保险公司
4	4	第四条　在保险年度内发生下列情况的，依照下列规定处理：　　（一）被保险人在我市范围内转学、转园的，可不办理保险关系转移手续，

Datensatz: 1 von 93024　Kein Filter　Suchen

6. Summary

This technical paper provided an introduction to the ongoing document analysis work in project B05 of the CRC 1342 in Bremen. It summarized the processes of data collection and setting up a document database, and subsequently presented two types of workflows used to leverage the data in ongoing publication projects. The first type relies on SQL queries and the identification of relevant policy keywords, and was used in a forthcoming publication in Asian Development Review. The second type relies on machine learning and text classification, and is used for ongoing publication projects on unemployment and health insurance.

These workflows provide but a glimpse of what the general approach has to offer. It can greatly facilitate the targeted analysis of large numbers of documents. Where keywords suffice for determining relevance, targeted SQL queries can identify relevant documents and sentences. Where more complex information needs to be identified, text classification via machine learning helps identifying such content. The results can be analyzed in both, qualitative and quantitative ways. In small or incomplete data sets, (like unemployment insurance or county-level documents), they can provide a quick overview of the diversity of local regulations, and help the researcher find the most relevant documents for a given question. In large and comprehensive datasets (like health insurance), it can additionally aid the creation of variables for quantitative analysis.

References

Duckett, Jane. 2011. *The Chinese State's Retreat from Health*. Abingdon: Routledge.

Frazier, Mark W. 2010. *Socialist Insecurity*. Ithaca: Cornell University Press.

Heilmann, Sebastian. 2008. "Policy Experimentation in China's Economic Rise." *Studies in Comparative International Development* 43 (1): 1–26.

Liu, Tao. 2015. *Globale Wissensdiffusion in Der Politik Sozialer Sicherung (Global Diffusion of Knowledge in Social Security Policy)*. Frankfurt am Main: Peter Lang.

Müller, Armin. 2016. *China's New Public Health Insurance*. New York: Routledge.

——. 2017. "Functional Integration of China's Social Protection: Recent and Long-Term Trends of Institutional Change in Health and Pension Insurance." *Asian Survey 57* (6): 1110–43.

Müller, Armin, and Tobias ten Brink. 2022. "The Diffusion of International Models in China's Urban Employees' Social Insurance." *Global Social Policy* 22 (3): 560–79.

Solinger, Dorothy. 2005. "Path Dependency Reexamined." *Comparative Politics* 38 (1): 83–101.

Wong, Christine. 2009. "Rebuilding Government for the 21st Century." *The China Quarterly* 200 (January):929–52.

World Bank. 2002. "China - National Development and Sub-National Finance." January 1, 2002. https://documents1.worldbank.org/curated/en/111911468240901599/pdf/multi0page.pdf