Dutch intellectual culture between 1962 and 1995, or, using classical algorithms and LLMs to efficiently extract data with imperfect OCR

Joris van Eijnatten

Introduction

Book reviews in newspapers provide a fascinating insight into past mentalities. Which books did book reviewers / journalists expect their readers to read, and how did this change over time? To reconstruct the 'intellectual culture' of an age, newspapers, given their past role as regular text-based information providers, are a highly apposite historical source. Relatively little historical work on book reviews has been done, and to my knowledge digital methods have never been applied to study them.

The use of digital methods is the second theme in this paper, which is part of a larger project that aims to convince a bigger cross-section of humanities disciplines to pay more serious attention to 'the digital'. Concretely, this paper experiments with ways to extract data from data, in this case to obtain book reviews from semi-structured data and identify the titles mentioned in those reviews. It employs a mix of classical algorithms, large language models and manual correction to (a) monitor the efficiency of the methods used and (b) construct a curated dataset from OCR'ed material of deficient and variable quality.

The dataset

The research makes use of the *Leeuwarder Courant (LC)*. This Dutch newspaper serves mainly as an example of the way a specific problem (sc. extracting book reviews) can be dealt with using familiar and easily accessible machine readable material. There are several advantages connected to the *LC* in particular. The newspaper is available in its entirety from the Second World War until 1995 (there are some gaps); this allows us to delve into a period that began and ended with two very different political, social and cultural ruptures: the War and the rise of the Internet. Although a regional medium, the newspaper enjoyed national prominence for the larger part of the period. Finally, its machine readability, while far from perfect, is relatively less bad than in other digitized newspapers.

Aim and method

The aim of this paper is twofold: to extract and identify book titles from newspapers, which can be seen as a form of data processing, and to analyze the resulting dataset with the actual research question in mind. Which books did the *LC* review between 1962 and 1995, and how did this reflect an evolving intellectual culture? This paper demonstrates various results: methods to extract book titles from imperfect data; an indication of the efficiency of the various methods employed; a structured and enriched database of book titles representing postwar Dutch intellectual culture; and an initial analysis of the results.

Skipping initial parts of the workflow, such as scraping the dataset from the National Library's API , the method of extracting the book titles from the newspaper consists of three main parts.

1. Classical algorithm.
Crucial for the workflow here is to identify an element (token) that occurs consistently in each book review and always occurs in a book title. We then extract 600 characters from the review that we can safely assume to cover (a substantial part of) a book title; we call this the 'title pericope'. Subsequently, we relate each title pericope to an external database of book titles. For this purpose

we used the ‘Nederlandse Bibliografie Totaal’ (NBT), an RDF-based database ideally covering all book titles published in the Netherlands. The trick is to determine which words the title pericope has in common with each book title occurring in the NBT, as the intersection between two sets. One way of doing this (several were used) is to state that if the intersection has x or more words in common, we assume that there is a match and that the title in the title pericope has, therefore, been identified.

2. LLM

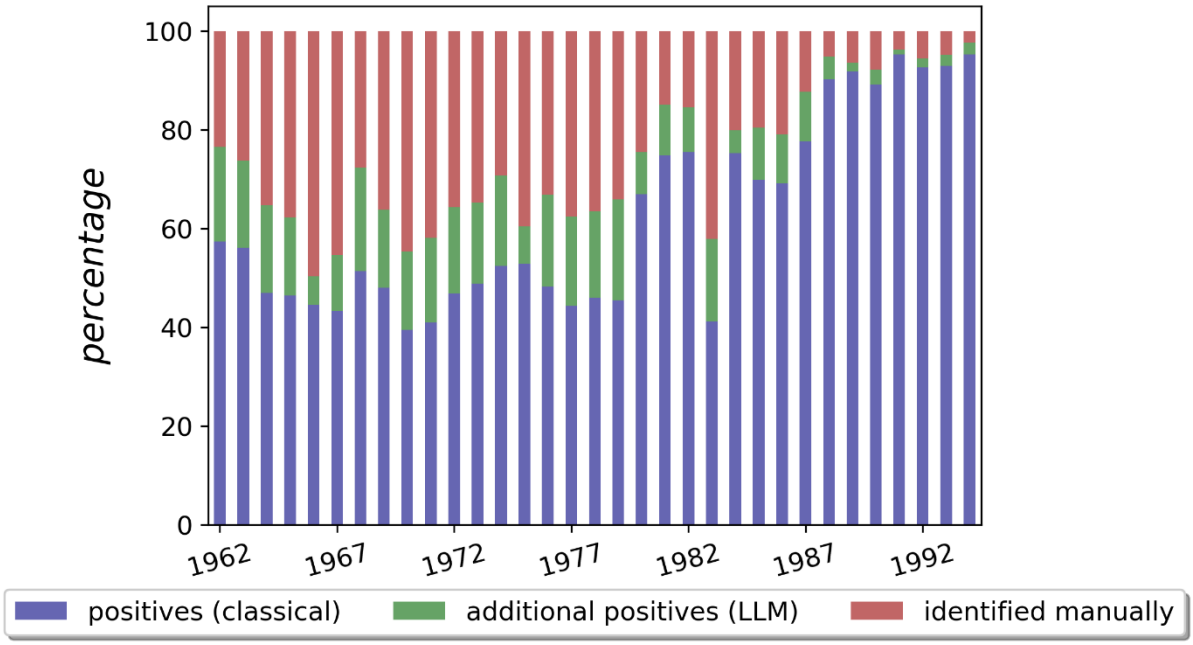
The results after applying the classical algorithm includes false negatives, in the sense that not all titles were able to be identified, partly due to faulty OCR. All title pericopes were therefore fed into an existing Large Language Model (in our case mainly ChatGPT 4, through an API) and the outcome was again related to the titles in the NBT. This improved the score.

3. Manual correction

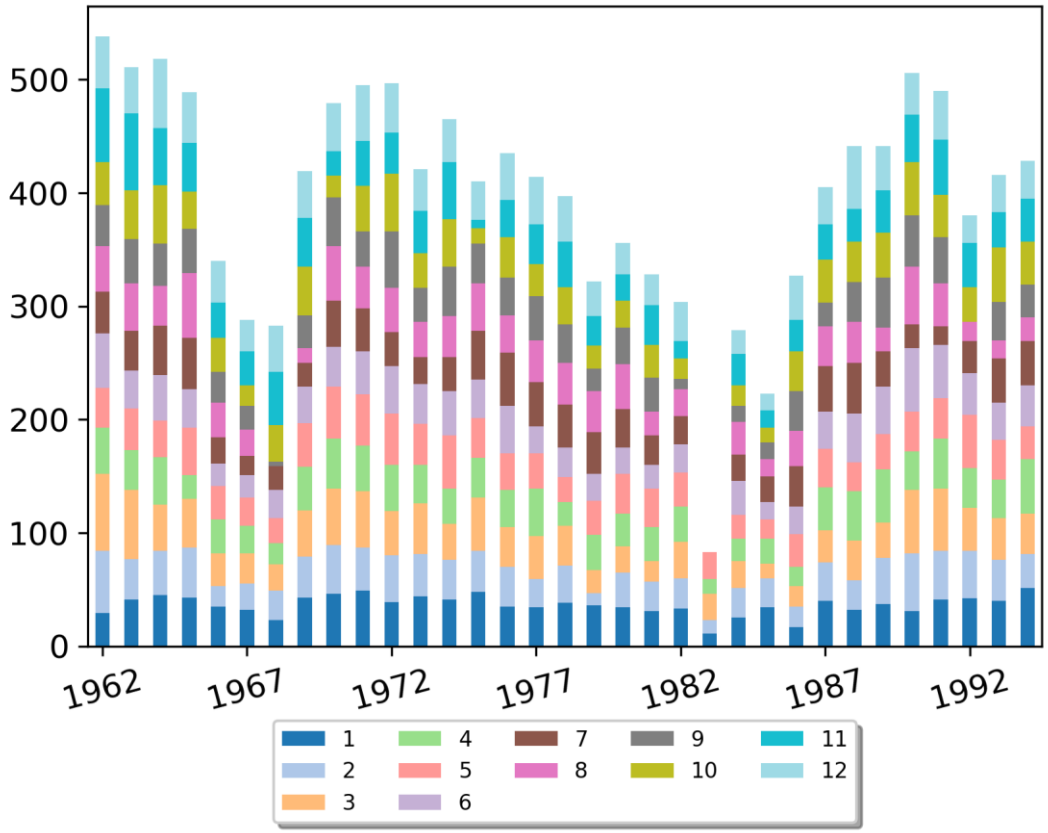
To determine the actual efficiency of the automated procedure, the whole dataset was corrected manually. This involved (a) checking the correctness of the title identification; (b) checking for false positives as well as false negatives.

Results

The resulting database, which thus includes information on the success of each step in the automated data extraction, as well as the actual book titles reviewed over the period under consideration, offers a wealth of data to use for a digitally informed historical narrative. The visualizations below offer some insights into the material.



Graph 1. The percentage of book titles identified using the three methods. The total number of titles is 25,940. The lower scores prior to the late 1980 are the result of a decrease in OCR quality but there are other variable, too. Interestingly, the 1970 saw a substantial increase in reviews of books published in Germany (which are not included in NBT).



Graph 2. The spread of review articles over months of the year (1-12).

name	no. of reviews
Simon Vestdijk	71
Georges Simenon	48
Theun de Vries	39
L. van Egeraat	38
Louis Couperus	36
Louis Paul Boon	35
Edgar Wallace	35
Johan Fabricius	35
Marnix Gijsen	34
Ellery Queen	32

name	no. of reviews
Agatha Christie	32
Hugo Claus	32
J. Bernlef	30
Harry Mulisch	29
Jan van Rheenen	29
Marten Toonder	29
Godfried Bomans	28
Bertus Aafjes	27
L. de Jong	27
Cees Buddingh'	27

Table 1. Top 20 of authors reviewed in *LC* between 1962 and 1994.

rank	name	frequency
1	Hitler	672
2	Jezus	480
3	God	337
4	Stalin	218
5	Marx	192
6	Kennedy	187
7	Napoleon	160
8	Lenin	135
9	Wilhelmina	129
10	Beatrix	118

Table 2. Top 10 of person names mentioned in reviews in *LC* between 1962 and 1994.

Reflection

One conclusion is that, at this juncture in time, a combination of methods is most fruitful, and that both are relatively successful in dealing with faulty OCR. A next step is to test other methods, such as Named Entity Recognition on titles and training BERT models.

Brief bibliography

Data:

- the *LC* is accessible through an API at the National Library of the Netherlands (kb.nl)
- the NBT was provided in its totality as an RDF database by the National Library of the Netherlands

Leeuwarder Courant:

- Marcel Broersma, *Beschaafde vooruitgang. de wereld van de Leeuwarder Courant 1752-2002* (Leeuwarden, 2002)

Book reviews and intellectual culture:

- Elizabeth Carolyn Miller, 'Reading in Review: The Victorian Book Review in the New Media Moment', in: *Victorian Periodicals Review* 49 (2016), 626-642
- Riie Heikkilä & Jukka Gronow, 'Stability and Change in the Style and Standards of European Newspapers' Arts Reviews, 1960–2010', in: *Journalism Practice* (2018), 12:5, 624-639, DOI: 10.1080/17512786.2017.1330664

Book reviews and DH:

- Matthew J. Lavin, 'Gender Dynamics and Critical Reception: A Study of Early 20th-century Book Reviews from The New York Times', in: *Journal of Cultural Analytics* (2020), 1-33 / DOI: 10.22148/001c.11831

LLMs

- Chat GPT3.5 turbo and Chat GPT4