Quantifying "Pillarization": Extracting Political History from Large Databases of Digitized Media Collections

Patrick Bos[1], Huub Wijfjes[23], Maaike Piscaer[2], and Gerrit Voerman[34]

[1] Netherlands eScience Center, Amsterdam, Netherlands,
p.bos@esciencecenter.nl
[2] University of Amsterdam, Netherlands
[3] University of Groningen, Netherlands
[4] Documentatiecentrum Nederlandse Politieke Partijen, Groningen, Netherlands

Abstract

We analyzed long-term dynamic developments in newspaper content in connection with the process of pillarization (the segmentation of Dutch society and politics along religious/ideological cleavages) over the period 1918–1967. One of the main characteristics of the historical debate on this phenomenon is an alleged close connection of political and media organizations on personnel, organizational and content-driven levels. In the political communication debate, this raises the question about the relationship between 'politization of media' and 'mediatization of politics'. Our preliminary study shows how patterns in the interactive relation between politics and newspapers can be further unraveled, by analyzing data based on the digitized newspaper collection of the Royal Dutch Library and related digitized political historical sources (parliament, parties, biographical material). In particular, it shows differences between Socialist and Catholic approaches towards the pillarized culture of "the other" and themselves. Whereas the Socialist pillar mainly focused on the politics and socio-economics of both their own and other groups, we find that the Catholic pillar was more inclined towards cultural and organizational self-reference.

1 Introduction

In this study, we combine analyses of digitized historical newspapers and political sources to reconstruct long-term patterns in a process called 'mediatization of politics in history' or, in specific Dutch context, pillarization of media and politics [1, 2, 3, 4]. Dutch society in much of the 20th century was vertically segregated into four main 'pillars', ideologically coherent groups: Catholics, Protestants, Socialists and Liberals. Each pillar had its own political parties, media outlets and other types of societal organizations. One of the defining properties of pillars was their strong internal interaction, for instance between parties and media [1]. This phenomenon has been extensively studied in the traditional historical context. We extend its study into the quantitative domain.

Digital media historical research is a challenging new specialization in the Digital Humanities [5, 6, 7, 8]. Research in digital newspaper archives — applying long-term analyses based on specific digital tools in combination with hermeneutic source critique — is in the early stages of development [9]. In our preliminary study into the historical context of pillarization of politics and media, we explore the possibilities and limitations of digital analysis of large databases.

To quantify the connection between media and politics, we designed four types of *indicators of pillarization* (Sections 3 and 4). These should help us understand texts by placing them in the right context in an automated manner. We could form a better understanding of current political discourse, and potentially identify emerging segregation / pillarization, or inversely, detect further globalization (which can be seen as the breakdown of segregation along nationality) of media. The indicators we developed are based on yearly counts of newspaper articles that contain pillar-bound words, like party names or ideological concepts. We tested our indicators on newspapers with diverse ideological backgrounds,

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

expecting a stronger "signal" from indicators associated with a certain pillar in the corresponding ideologically affiliated newspapers than in newspapers of other pillars.

The two periods of interest in this work are the "interwar" years of 1918–1940 and the "reconstruction" period of 1946–1967.[1] This timespan saw the peak of pillarization before World War II and the beginning of "depillarization" at the end of the sixties. Given the restrictions in the digital availability of relevant newspapers (Section 2), we focus mainly on Catholic and Socialist groups and their newspapers. For these two pillars, the most complete newspaper collection is available between 1918–1967.

2 Data Description and Management

To study pillarization in connection to media content, we set out to obtain a large, unified database of Dutch newspapers. The Royal Dutch Library (*Koninklijke Bibliotheek*, KB) has digitized (scanned and OCR'ed) a large collection of both national and regional newspapers published between 1618 and 1995. Their complete database is freely available for scholarly purposes. This forms the basis of our dataset. Additionally, we obtained data on political parties, leaders, communities and ideological concepts from:

- digitized biographical sources from the *Parlementair Documentatie Centrum* at Leiden University;
- digitized Proceedings of Dutch parliament from the PoliticalMashup project [10];
- digitized party programs or declarations of principles of political parties.[2]

In what follows, we elaborate on the newspaper dataset and highlight some challenges we encountered.

2.1 Newspaper Data Format, Metadata and Completeness

Apart from having OCR'ed the newspaper texts, the KB has subdivided newspapers into separate articles. This guarantees a basic level of topical unity that greatly aids us in our analysis. The rest of the page or column (a form of 'noise') can be cleanly separated from the relevant article (the 'signal'), increasing the relevance of search results. Other relevant metadata fields are newspaper title, newspaper date and article title; many other metadata fields are present, but we did not use these.

We set up a (partial) copy of the full KB database (Section 2.2), without further modifications. We enriched the article data by adding — based on our own knowledge of historical literature — the ideological context (pillar or neutral) of the article's newspaper.

An overview of our dataset of newspaper articles is given in figures 1 and 2. The KB collection does not yet offer complete coverage of all relevant newspapers (for coverage evaluation see [9]). Overall, the completeness of the corpus is greater in the interwar years than in the later reconstruction period.

The Socialist collection contains almost all significant titles of the total Socialist press and is therefore the most complete and representative part of the corpus. In the Catholic collection, some important titles are missing, like "De Gelderlander", "Brabants Dagblad" and other large (southern) regional papers. The neutral group (titles not affiliated with the main pillars) is also relatively incomplete, with many regional and city-based titles with high circulation figures missing. In the interwar years, the availability of the Liberal collection is scattered, while after 1945 Liberal papers are completely lacking, even though Liberal papers had high circulation numbers and were widely considered important for political orientation, due to their strong focus on political reporting. Given that the Protestant pillar was one of the leading groups in Dutch society, the complete lack of Protestant newspapers created an important restriction to this study. This situation led us to focus our analysis on the Catholic and Socialist pillars that provided the most complete sets of digitized newspapers in the period 1918–1967.

[1] Specifically, we define our periods between election dates; the "interwar" period is set between 1918–07–03 and 1940–05–10 and our "reconstruction" period takes place between 1946–05–16 and 1967–02–15.

[2] "Beginselprogramma's" from the DNPP repository, accessible through `http://dnpp.nl/themas/beginsel`.

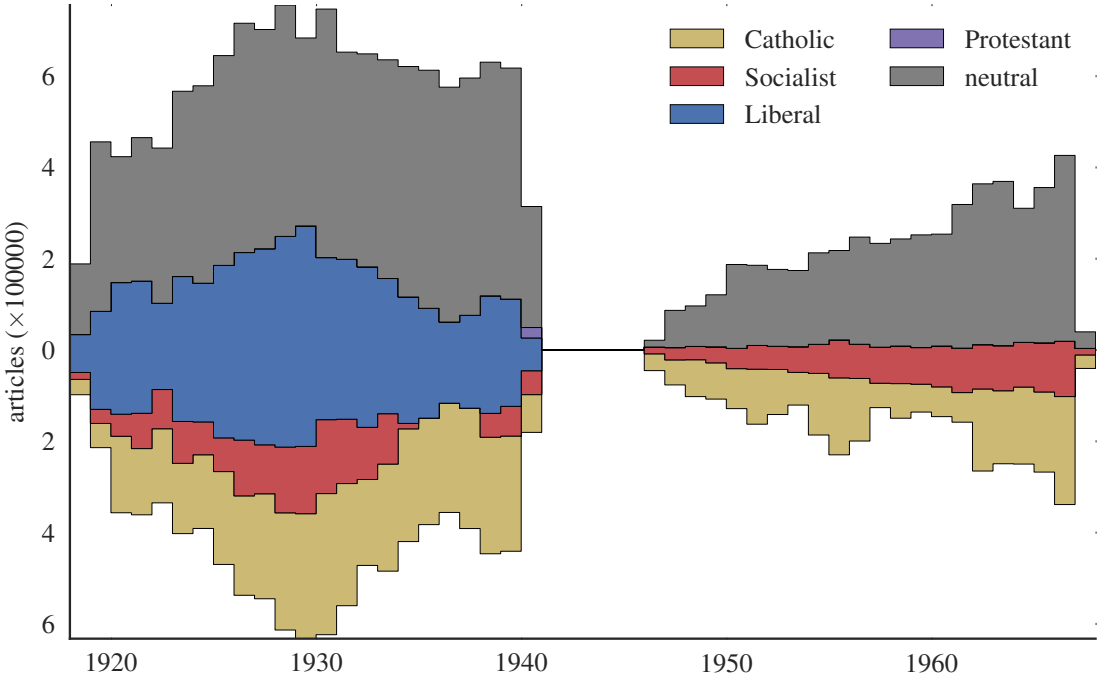


Figure 1: Number of articles per year for each pillar in the KB newspaper corpus used in this study.

2.2 Newspaper Database and Search Engine

To store and query our newspaper articles and their metadata we use Elasticsearch (ES), a document store database and search engine. It can easily process large text-based datasets by making efficient use of distributed systems.[3] ES is optimized for analytic queries (filters and aggregations) and ranks among the fastest full-text search engines currently available. This enables us to do real-time analysis on the large dataset, allowing for a quick feedback cycle in the development process.

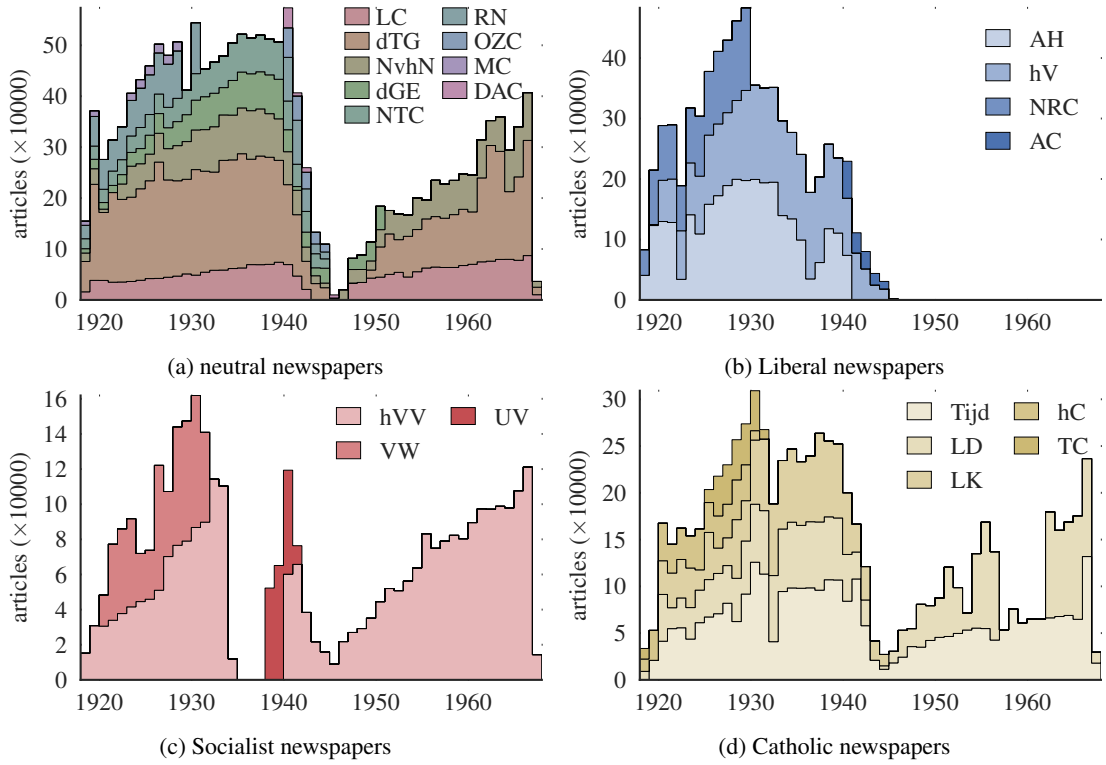
To enable fast searches, ES builds an inverted index of the words in all article texts. The index links a term to all the articles where it occurs, as well as the term's location in each article. Before words can be indexed, the raw article texts must be "analyzed", i.e. processed into index terms. In our case, we used the ES built-in Dutch Analyzer for this. Article texts are tokenized (splitting of sentences into words, also removing punctuation), turned into lower case to make search case-insensitive, filtered for Dutch stop words[4] and stemmed[5] (e.g. turn "working", "worked" and "worker" into "work").

When searching for a specific query in the database, the query text is analyzed in the same way as the article texts. This allows ES to match to index *terms* instead of literal words. This way, when searching e.g. for "socialist", the engine can also be made to search for "social" (depending on the language and exact implementation of the Analyzer).

The indexing procedure further produces a term vector for each article, which counts the index terms in the article. These are used to quickly calculate statistics of selected articles called "aggregations".

ES offers many possible types of queries and filters. For our analysis, we used the "query string" query type together with filtering by year and pillar. The "query string query" is a boolean query type

[3] Our ES (https://www.elastic.co/products/elasticsearch) setup runs on 6 machines maintained by SURFsara.
[4] List of Dutch stop words can be found at http://snowballstem.org/algorithms/dutch/stop.txt.
[5] According to a Snowball stemming algorithm (http://snowballstem.org/algorithms/dutch/stemmer.html).



(a) neutral newspapers

(b) Liberal newspapers

(c) Socialist newspapers

(d) Catholic newspapers

Figure 2: KB newspaper articles per year and pillar, subdivided by each pillar's newspapers. Newspaper title acronyms: *neutral:* LC: Leeuwarder Courant, dTG: De Telegraaf, NvhN: Nieuwsblad van het Noorden, dGE: De Gooi- en Eemlander, NTC: Nieuwe Tilburgsche Courant, RN: Rotterdamsch nieuwsblad, OZC: Overijsselsche en Zwolsche courant, MC: Middelburgsche courant, DAC: Drentsche en Asser courant. *Socialist:* hVV: Het (vrije) Volk, VW: Voorwaarts, UV: Utrechts volksblad. *Liberal:* AH: Algemeen Handelsblad, hV: Het Vaderland, NRC: Nieuwe Rotterdamsche Courant, AC: Arnhemsche courant. *Catholic:* Tijd: De Tijd, LD: Limburgsch dagblad, LK : Limburger koerier, hC: Het Centrum, TC: Tilburgsche courant.

that can search for specific combinations of multiple terms by using "AND"s, "OR"s and parentheses. It allows us to search simultaneously for several multi-word variations of indicators of pillarization (see Section 3). This reduces the number of queries significantly, since most queries contain many variants.

3 Measuring Quantitative Indicators of Pillarization

We search our dataset of digitized, OCR'ed newspaper articles for terms related to specific pillars. The four indicators we define are:

- Reference to political parties (names, acronyms);
- Reference to party first candidates in election times (names);
- Reference to non-political organizations (names and acronyms): cultural (excluding media), religious, societal (housing, health care), social-economic (trade unions, employers, professional organizations), educational (schools, universities);
- Reference to ideologically charged concepts (see Section 4).

We quantify these indicators by searching the dataset for articles with these references and counting the articles. The articles are counted per year. Also, we cluster newspapers by their affinity to a certain

pillar (or to a neutral position) and compare the counts of indicators in the different pillar clusters. The article counts for each of the indicator categories are evaluated using the Kullback-Leibler divergence with respect to the total corpus. In this section, we explain how this measure relates to raw article counts.

3.1 Article Count and Frequency

The basic unit of measurement from our database searches is the raw number count of articles $N_t(c,p)$ that are found to contain an indicator term t (or derived, "analyzed", term thereof, see Section 2.2). We bin the counts by cluster c and by period p. Cluster c might be a group of newspapers belonging to a certain pillar or even the entire collection and the period p is one year.

To compare the indicator counts in different clusters (pillars, etc.), we need to take into account that the total number of articles differs for each cluster. To this end we define the indicator frequency f:

$$f(c,p) = \frac{N_t(c,p)}{T(c,p)}\,, \tag{1}$$

where $T(c,p)$ is the total number of articles in cluster c in period p. The frequency measures the percentage of articles that contain indicator t, per cluster and period.

It is important to note that when the total number of articles in a cluster is small, one must take care in the interpretation of the frequency. A small number of articles might give a less balanced, more biased view of the opinions within a cluster.

3.2 Specific Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence D_{KL} is a measure of the difference of one distribution as compared to another. In general, it is defined as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}\,, \tag{2}$$

where P and Q are distributions over some variable i.

We are interested in the difference of indicator frequency distributions *over time*. Were we to use the KL-divergence strictly as is, we would lose our temporal dimension, since all variables of the distributions must be summed over. We use a slightly modified measure for our analysis, a "specific" Kullback-Leibler (sKL) divergence. It is given for each period p (year) as:

$$D_{\mathrm{sKL}}(p)(c \parallel d) = f(c,p) \ln \frac{f(c,p)}{f(d,p)}\,, \tag{3}$$

where we compare the indicator frequencies of the pillarized clusters c to those of the total database d, i.e. all newspapers regardless of pillar. Compared to the frequency, the sKL-divergence more cleanly separates general trends in media coverage from pillar-specific trends.

4 Ideologically Charged Concepts

One of the key indicators offered the greatest challenge for digital operationalization. The "ideologically charged concepts" are words or combinations of words that mark political or pillarized identity. Examples of such concepts are "klassenstrijd" (class struggle) or "socialisatie" (socialization) for Socialists, "vrije markt" (free market) for Liberals and "subsidiariteit" (subsidiarity) for Catholics. We extracted

these terms semi-automatically from the ideological party programs that represent the party's political ideals and basic principles in the relevant period.

The first step in our extraction procedure is to identify concepts automatically by looking for words in a document that particularly distinguish it from similar documents. For this, we use a so-called "parsimonious language model" to simultaneously filter out two sources of noise from the party programs:

1. Stop words and other common language;
2. Common political jargon.

To filter common political jargon out of party programs texts, we used a more or less politically neutral set of background documents: the digital proceedings of Dutch parliament.[6]

The parsimonious language model [11, 12] is a unigram model[7] that combines maximum likelihood estimates of the concept probabilities in a document (a party program) with those in the background corpus of documents. If the probability of a word in a document is high — not only in the document itself, but also as compared to the probability in the background corpus — the "parsimonious probability" of the word will be high as well. It will be low if the word has similar probabilities in the document and in the background corpus. It is likely to be a common word (or, in this case, common political jargon), which does not particularly distinguish the document. We extract the ideologically charged terms pertaining to the party and hence the pillar by calculating the parsimonious probabilities of the words in the program documents and considering only the 50 words with the highest probabilities.

One issue that should be carefully considered is how to select which proceedings to use as a background for each program. The programs strictly apply only to the period in which the program was written. Also, over the course of the period we study in this work, the Dutch language evolved — both in general, and in the political discourse. It therefore makes sense to make a time-bound selection of the parliamentary proceedings to match with the time frame of each program. We tested several period length selections of proceedings, from one up to ten years. We found that one year of proceedings was sufficient as a background; using more than that did not significantly alter the outcome of the model's probability distribution. We chose to only use the year before (not the one after) a program was published. This prevents that new language that was not yet there at the time of writing the program influences the model in some unexpected way.

In our work, the parsimonious language model is not calculated for each party program separately, but rather on an aggregation of all programs in each of the two periods (the interwar years and the reconstruction period). This way, we need not worry about the fact that the programs were generally published in different years for different parties. Moreover, clustering over a period offers the benefit of a long-term view, which might reveal insights in (the continuity of) parties' ideologies. For the scope of the study of pillarization, these periods identify the main levels of aggregation at which differences between these indicators would be expected. A related issue that is addressed by clustering over a period is that parties generally drafted their ideological programs in different years, contrary to their election programs. Strictly, therefore, one program cannot be cleanly compared to any other, since the political landscape might have changed significantly in the meantime. By averaging over the entire period, we gain the ability to compare parties in a methodologically sound manner.

Since we clustered the programs per period, we used the entire available set of proceedings over the two periods as backgrounds. As discussed above, this may introduce problems with language usage of the end of the period interfering in unexpected ways with the programs at the beginning of the program (and vice versa). To solve this, one might need to modify the language model to explicitly include a

[6] Of course this source is not perfectly neutral. The composition of parliament might affect matters, as well as the vocalness of certain members and events like elections or big societal events. However, the considerable scale of the Proceedings guarantees neutrality more than other documents.

[7] A unigram language model assigns probabilities $P(t|D)$ of words t occurring in documents D based on some statistics of a corpus, for instance the count of word t divided by the total number of words in the corpus.

time parameter to weigh the influence of documents in the background collection. This goes beyond the scope of this work.

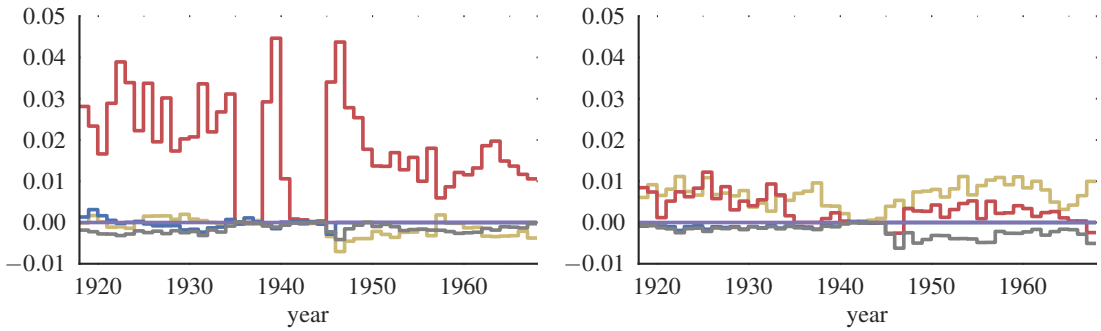
We checked the resulting lists by hand (using expert knowledge about historical contexts) to remove some noise (some of the documents contain URLs or uncommon abbreviations like "art." instead of "article") and in some cases combine the unigram terms into N-grams.

5 Results and Historical Context

In what follows, we present a selection of indicator query results, accompanied by a short interpretation within the historical context. Comparing the results of our searches to the historiography of pillarization, we can heuristically assess the sensitivity of the different indicators. Overall, we find that the indicators align well with what we already know about pillarization. This means that these indicators could subsequently be used in other studies as a means of quantifying the pillarized nature of a text.

The figures in this section show the sKL-divergences of the indicators from the total corpus. The same color scheme is used to indicate the different pillar groups in each figure: Catholic newspapers signals are yellow, Socialists red, Liberals blue, Protestants purple and neutral newspapers are gray. As mentioned before, we focus our analysis on the Socialist and Catholic signals.

In general, the intensity of indicators of a certain pillar within the cluster of that same pillar can be seen as a measure of "intrapillarization". It is assumed to probe the internal self-promotion of the pillar. This can be contrasted with "interpillarization", which concerns the interaction between pillars. In most cases this is expected to be of an antagonistic nature. The intensity of these two exponents of the general phenomenon of pillarization might differ per pillar. By studying these two measures separately, we gain insight on the nature of the pillar as it evolves over time (is it focused on itself or also on other pillars?), as well as a general measure of its "pillarizedness".

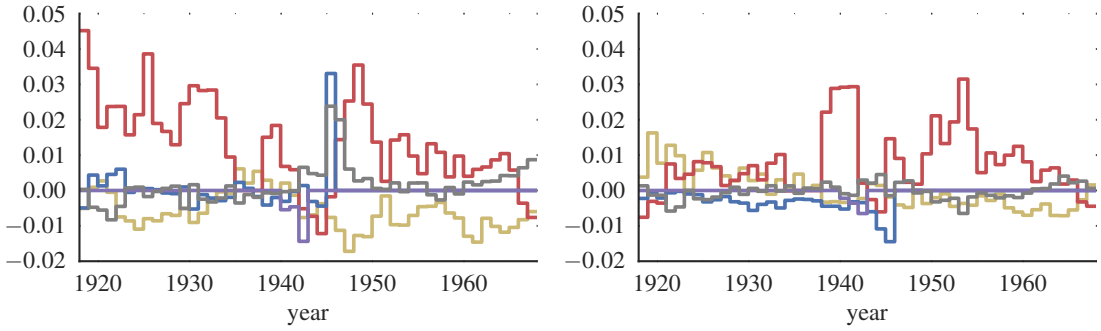


(a) Socialists (SDAP & PvdA)

(b) Catholics (RKSP, ABRKV & KVP)

Figure 3: Mentioning of party names ("signal") in newspapers affiliated to pillars, counted per year. Yellow: Catholic, red: Socialist, blue: Liberal, purple: Protestant, gray: neutral (see text on page 7).

Party Names. Figure 3 shows that the Socialist "signal" (mentioning of party names) is generally higher than the Catholic signal. The sKL-divergence is lower in the reconstruction period, especially for the Socialist indicators. The level of the Catholic signal remains stable over the entire period. Moreover, the Catholic newspapers refer to their own party significantly more than to the Socialist parties. The Socialist newspapers, in contrast, refer to all parties. This indicates that party politics is a dominant Socialist theme.

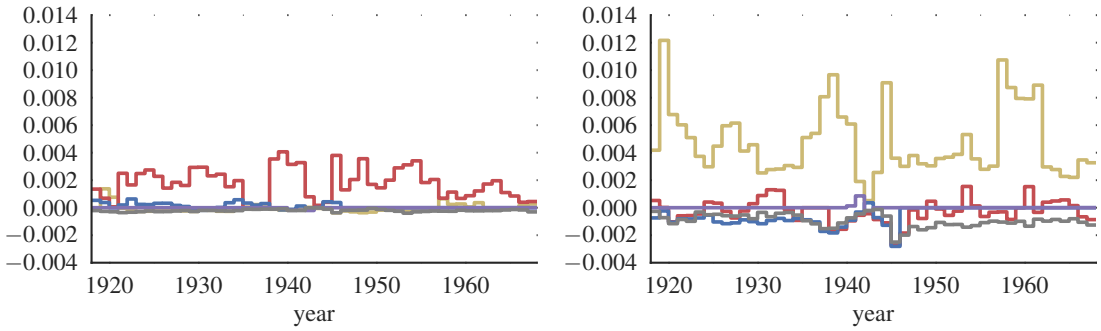


(a) Socialists (SDAP & PvdA)

(b) Catholics (RKSP, ABRKV & KVP)

Figure 4: Mentioning of party first candidates in newspapers affiliated to pillars, counted per year. Yellow: Catholic, red: Socialist, blue: Liberal, purple: Protestant, gray: neutral (see text on page 7).

Party First Candidates. In general, the patterns in the party first candidates signals are similar to those of the party names, as shown in figure 4, except for the Catholic pillar, for which first candidates do not seem to be a strong indicator. Again, the dominance of politics in the Socialist pillar is contrasted by the moderate political interest from the Catholic pillar. It seems that in election years (e.g. around 1925 and 1933) the signals are stronger, which is expected for the position of candidates in election times.

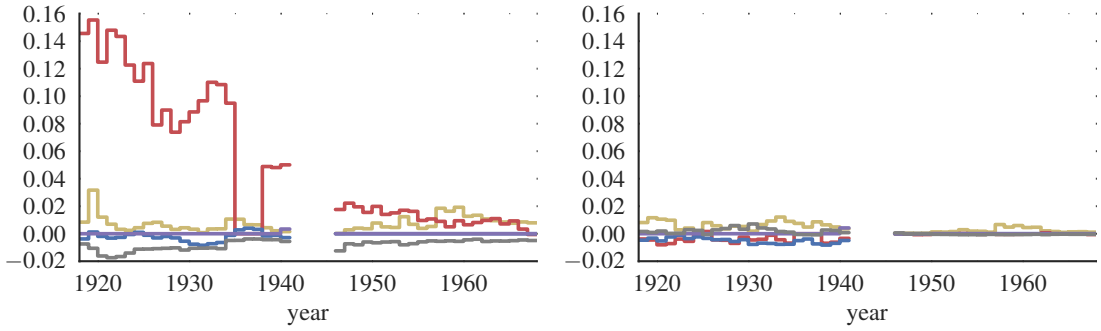


(a) Socialist organizations

(b) Catholic organizations

Figure 5: Mentioning of organizations in newspapers affiliated to pillars, counted per year. Yellow: Catholic, red: Socialist, blue: Liberal, purple: Protestant, gray: neutral (see text on page 7).

Organizations. Whereas the Socialist newspapers speak a lot about the Catholic political organizations, this is not the case for the non-political organizations of the "organizations"-indicator, as illustrated in figure 5. This indicates that the strong attention of the Socialists for the Catholics mainly had a political dimension. The Catholic pillar has a more prominent focus towards (their own) societal organizations. This highlights the difference in strategies of the Catholic and Socialist pillars; Catholics were community builders (internal approach), whereas Socialists focused on the political and socio-economic situation of the working class, also in other pillars (external approach). For the Catholics, pillarization is more a question of organizing their own societal and cultural architecture.



(a) Socialist terms

(b) Catholic terms

Figure 6: Mentioning of ideological terms in newspapers affiliated to pillars, counted per year (see p. 7).

Ideological Terms. The reference of the Socialist newspapers to their own conceptual context is strong in the interwar period, but steadily declines towards the period of depillarization. While the Socialist indicator in the Catholic newspapers remains more or less stable in the interwar period, it slowly rises in later years to around the same level as the Socialist newspapers. This is an indication of the rising need for dialog in the political coalition of Socialist and Catholic parties in the years after 1948.

Whereas Socialist newspapers did show strong signals for previous Catholic indicators, in this case, there is a markedly lower article count for Catholic ideological concepts. The underlying strategy behind this may be the notion that one cannot win Catholic voters for socialism by attacking or criticizing the Catholic values and concepts. The Catholic concepts are more or less exclusively used by Catholic newspapers. These Catholic concepts score significantly higher in the article count for Catholic newspapers. This confirms the hypothesis that Catholics are primarily self-referential and that their ideology is less accessible from a non-Catholic outsider perspective.

6 Discussion and Conclusions

The main picture that presents itself from our study is that long-term historical patterns can indeed be recovered in a quantitative way that matches well with our previous knowledge from the historical literature. Of course, such patterns do not tell us the full story of highly complex societal topics like pillarization, but dominant trends clearly emerge. With further study, a more detailed picture may be obtained, for instance by adding information from sentiment analysis of texts, which would tell us not only whether an article mentions an indicator, but also whether it is positively or negatively regarded. One would expect newspapers from one pillar to often speak negatively about other pillars, for instance, which we cannot infer from the indicators we presented here.

Having verified that societally and/or politically charged indicators can indeed be applied to media content, our indicators could subsequently be applied to other texts. This way, one could classify the ideological affiliation of texts or authors, or even complete newspapers or other works, in a way similar to sentiment analysis tools that detect emotionally charged words in texts. The raw lists of indicators are freely available online[8], as well as the code for using them to obtain the results in this work.[9]

The other main conclusion from our study concerns our general experiences with combining quantitative and hermeneutic methods. We identified three general issues that should be addressed in the near

[8] https://www.kb.nl/organisatie/kb-fellowship/huub-wijfjes

[9] https://bitbucket.org/egpbos/pidilib

future to boost humanities research by avoiding reinventing the wheel over and over again:

1. Incomplete coverage of newspaper collections hinders systematic studies like ours.
2. OCR data is still quite noisy, especially for short words and abbreviations. We dealt with this by omitting short words from our indicators, excluding many party name abbreviations.
3. Apart from the KB, a number of regional newspaper archives exist (in the Netherlands), as well as many other data providers. Most of these organizations offer access to their data for scholarly purposes. However, the logistics involved and digital expertise necessary for obtaining, combining and analyzing these data form a significant stumbling block for humanities scholars to enter the field of Digital Humanities. We set up our own search engine, but it would have been preferable to work with an existing database. Unfortunately a central, unified place for researchers to access and at the same time analyze the data does not yet exist.

References

[1] Arend Lijphart. *The politics of accommodation: Pluralism and democracy in the Netherlands*, volume 142. Univ of California Press, 1975.

[2] Klaus Arnold, Christoph Classen, Susanne Kinnebrock, Edgar Lersch, and Hans-Ulrich Wagner. Von der politisierung der medien zur medialisierung des politischen. *Zum Verhältnis von Medien, Öffentlichkeiten und Politik im*, 20, 2010.

[3] Huub Wijfjes and Gerrit Voerman. *Mediatization of politics in history*. Leuven: Peeters, 2009.

[4] Stig Hjarvard. *The mediatization of culture and society*. Routledge, 2013.

[5] Adrian Bingham. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History*, 21(2):225–231, 2010.

[6] Bob Nicholson. The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History*, 19(1):59–73, 2013.

[7] Joris Van Eijnatten, Toine Pieters, and Jaap Verheul. Big data for global history: The transformative promise of digital humanities. *BMGN-Low Countries Historical Review*, 128(4), 2013.

[8] Marnix Beyen. A higher form of hermeneutics?: The digital humanities in political historiography. *BMGN-Low Countries Historical Review*, 128(4), 2013.

[9] Huub Wijfjes. Digital humanities and media history: A challenge for historical newspaper research. *TMG-Journal for Media History*, forthcoming, 2016.

[10] Maarten Marx, Nelleke Aders, and Anne Schuth. Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, pages 99–104. Digital Government Society of North America, 2010.

[11] Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 178–185, New York, NY, USA, 2004. ACM.

[12] Rianne Kaptein and Maarten Marx. Focused retrieval and result aggregation with political data. *Information Retrieval*, 13(5):412–433, 2010.