Melvin Wevers*

Mining Historical Advertisements in Digitized Newspapers

https://doi.org/..., Received ...; accepted ...

Abstract: Historians have turned their focus to newspaper articles as a proxy of public discourse, while advertisements remain an understudied source of digitized information. This paper focuses on how historians can use computational to work with extensive collections of advertisements. First, we can use analyze metadata to understand better the different types of advertisements, which come in a wide range of shapes and sizes. Information on the size and position of advertisements can help to construct particular subsets of advertisements. Second, this chapter describes how we can extract textual information from historical advertisements that can be used for historical analysis of trends and particularities. For this purpose, we present a case study based on cigarette advertisements.

Keywords: historical advertisements, text mining, digitized newspapers, digital history

1 Introduction

In recent years, we have seen an explosive growth of digitized historical newspapers. Innovations in natural language processing have extended the possibilities for historians to extract information from large corpora of digitized historical texts. One of the sources that has been digitized extensively is newspapers. National libraries and project such as Impresso, Newseye, and Oceanic Exchanges offer access to digitized archives of historical newspapers.[1]

For many historians, newspapers provide a longitudinal understanding of public discourse [Postman, 2005, van Vree, 1989]. Ostensibly, newspapers are not the only gateway to public discourse, since they do not capture public discourse in its entirety. In its function as a proxy, a newspaper operates as a transceiver; it is both the producer and the messenger of public discourse [Schudson, 1982, 17-8]. On a surface level, newspapers inform us about the views of journalists and people

1 https://impresso-project.ch/, https://www.newseye.eu/, https://oceanicexchanges.org/

***Corresponding author: Melvin Wevers,** DHLAB - KNAW Humanities Cluster

that were interviewed by these journalists. However, as Margaret Marshall claims, scholars can also uncover the "values, assumptions, and concerns, and ways of thinking that were a part of the public discourse of that time" by analyzing "the arguments, language, the discourse practices that inhabit the pages of public magazines, newspapers, and early professional journals [Marshall, 1995, 8]." With the easier access to and increased availability of newspaper repositories, studies into the representation of ideas, values, and practices in public discourse gained traction [van Eijnatten and Ros, 2019, Daems et al., 2019].

Even though newspapers contain a considerable amount of advertisements, these remain an understudied source in computational studies of public discourse. This is surprising since advertisements are rich and varied carriers of information of the past. In his seminal work *Advertising the American Dream*, Roland Marchand argues that adverts provide a lens on the past. He writes that ads provide "insight into the ideals and aspirations of past realities [...] they show the state of technology, the social functions of products, and provide information on the society in which a product was sold [Marchand, 1985]." Others point out that even though adverts provide perspectives on the past, this is a distorted one, as the content of ads is driven by commercial interest [Fox, 1997]. Marchand acknowledges this criticism and conceptualizes advertisements as distorted mirrors. He argues that despite the primary function of advertisements to sell products, they still communicated social and cultural values, albeit in a somewhat distorted manner [Marchand, 1985, Lears, 1994].

Moreover, one could argue that for ads to be succesfull, they needed to resonate with their audience. As such, they had to be reflective of aspects of public discourse. Still, it is crucial to remain critical of the skewed representation of ideas and values in historical advertisements.

Advertisements are a fascinating and complex historical source, partly because of their multi-modal nature; they contain both visual and textual content. Since there often is an interplay between the visual and textual material, an analysis of only the textual content is somewhat limited. More recently, advances in computer vision provide methods to examine digitized visual material at scale. While the use of computer vision is rapidly evolving and already offers promising methods of analysis, it falls outside of this chapter's focus, which is on the analysis of metadata and textual content [Wevers and Smits, 2020, Arnold and Tilton, 2019, Chung et al., 2015].

The ability to extract information from large numbers of advertisements, both synchronically and diachronically, allows scholars to study cultural expressions on a macro-scale. However. it also enables fine-grained contextualizations of specific expressions. More often than not, historical interpretations drawn from advertisements are based on a small selection of ads, which might lead to cherry-

picking. The ability to chart trends over time makes it possible to model whether findings in smaller subsets can be generalized, or whether specific expressions deviate from general patterns. One an intermediate, mesoscale, we can also more easily find variations of a single cultural expression. These insights gained with computational methods supplement existing modes of analysis.

At the same time, we have to take into account that the quality of digitized historical sources is often sub-optimal, despite advances in Optical Character Recognition (OCR) software and natural language processing. It is especially challenging for OCR software to recognize text in advertisements correctly. In addition to more generic factors such as the quality of paper and or the printing technique, text in advertisements is often presented in varying sizes and fonts or as part of a heavily-stylized logo. OCR software regularly turns these forms of textual content into gibberish. The suboptimal text quality makes it more challenging to study adverts than articles.

Despite these shortcomings in OCR software, we can still extract meaningful information from the OCR-ed text. Given the amount of data, we can still find patterns in advertising discourse despite the noisy nature of the text. Although we have to be selective of the methods that we use, since their performance is impacted in different ways by imperfect OCR [van Strien et al., 2020]. In addition to the OCR-ed text, we can also learn about trends in advertising from metadata on the position and size of advertisements in newspapers. We can, thus, study multiple aspects of advertisements using computational means.

This chapter showcases how we can use computational techniques to study advertisements at scale in extensive collections of digitized newspapers. The first section shows how we can use metadata to examine trends in advertising. The second section gives examples of how text mining can be used to extract information from advertisements. This step is explained by demonstrating a case on product nationalities associated with cigarettes. In this case study, we rely on text mining to better understand changes and continuities in the associations with nationalities in advertisements for cigarettes.

2 Metadata Analysis: Advertisements in Almost All Shapes and Sizes

The digitization of newspapers has made it possible to study advertisements at scale using key word searches. However, not all digitized newspaper collections have segmented the articles and advertisements, presenting users with full-page scans that multiple document types. Recently, there have been efforts to rely on

computer vision to segment and classify document types in newspapers.[2] In cases where text blocks were segmented but not classified, text classifiers can be used to identify document types [Bilgin et al., 2018]. After training a text classifier with annotated data, the classifier could then predict metadata for unseen new data.

Fortunately, Delpher, the digitized Dutch newspaper archive hosted by the National Library of the Netherlands (KB), includes segmented documents and metadata on the document type.[3] Access to this type of metadata allows researchers to filter for advertisements. For the twentieth century alone, we can quickly assess that Delpher holds over thirty million advertisements in national and regional newspapers. Browsing through the results, it becomes clear that advertisements come in a wide range of shapes and sizes. We can find tiny, square-shaped advertisements; column-shaped classifieds advertisements; and full-age spreads (see Figure 1 for an example of a column of classified ads). The content and target audiences of these ads vary considerably. How can we deal with this variation, without more specific metadata information? This section shows how we can use metadata on the size and position of advertisements to cluster types of advertisements and subsequently filter out subsets. In what follows, we focus on two national newspapers to show how metadata can help to create subsets of advertisements but also provide insight into advertising practices.

Fig. 1: Classified adverts from *De Tijd*, 19 January, 1957

The analysis in this section is based on the metadata of the national newspaper *Trouw* (1946-1995).[4] An initial exploratory data analysis shows that the dataset contains instances of bad segmentation, in which small parts of advertisements appeared as separate advertisements. To exclude these segmentation errors from the dataset, we filtered out ads with a width or height smaller than 100 pixels. After removing these advertisements, *Trouw* contains about 1.31 million advertisements for the period 1945-1995. From Figure 2, we can gauge that the total number of advertisements increased between 1946 and 1995, with a sudden increase in the early 1980s. Alongside this increase in the number of ads, we also see a sudden, albeit slight decrease in the size of

2 See, for example, Ben Lee's project *Newspaper Navigator* at the Library of Congress. https://github.com/LibraryOfCongress/newspaper-navigator

3 http://www.delpher.nl. The metadata on document type was manually added.

4 For most of the examples in this chapter, direct access to the metadata and newspaper data is required. Researchers can contact KB for API access to the newspapers.

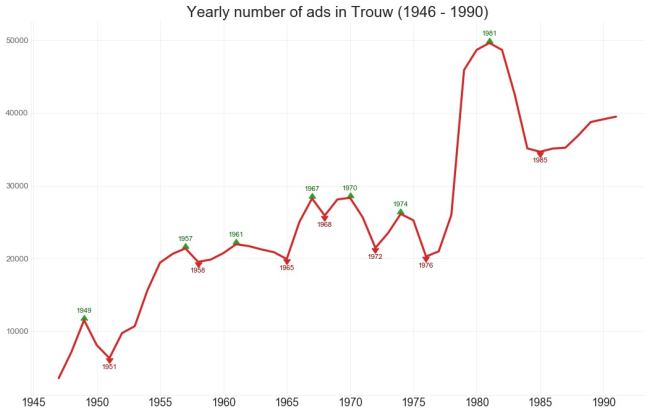


Fig. 2: Yearly number of advertisements in *Trouw*. Green triangles indicate peaks and red triangle troughs.

the advertisements (Figure 3). We plotted took the log_{10} of the mean size to visualize the rate of change better.

In addition to their size, we can also examine where ads are positioned in the newspaper and where on the individual pages. Using the width and height information of advertisements, we can create a heatmap that shows which parts of the page are taken up by advertisements. There is a clear difference in pixel density between odd and even pages in *Trouw* (Figure 4a & 4b). This can be explained by the fact that advertising on odd pages is more expensive than on even pages, as well as the common practice that new sections generally start on the odd pages. Moreover, we see that on the even pages, the upper-left corner, which is closer to the margin, is less populated than the lower-right side.

To get a better grasp of the different clusters of ads based on their size and their position in the newspaper, we can also apply unsupervised clustering methods. We base this clustering on the features: width, height, and relative position of the ad in the paper. The latter indicates how close the ad is to the front page or the last page. The distribution of these features contains multiple peaks, indicative of multiple sub-distributions. To be able to capture the distributions of these subpopulations, we can use Gaussian Mixture modeling, which can estimate

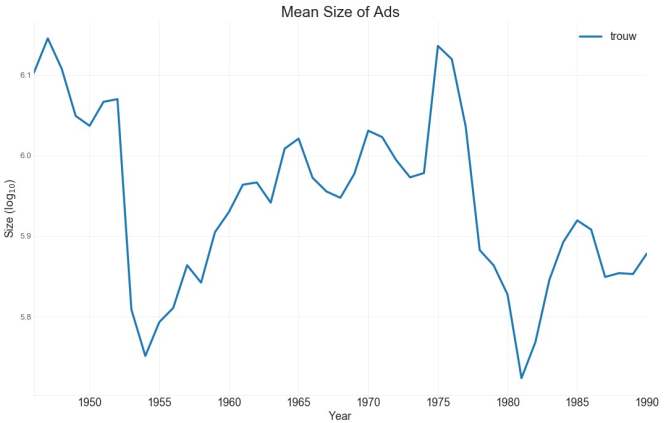
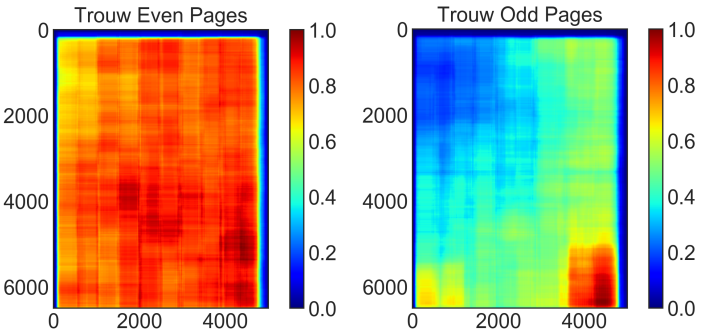


Fig. 3: Mean size of advertisements in *Trouw*

the parameters of these mixtures.[5] Compared to K-means clustering, mixture models incorporate information on the covariance structure of the data, making it possible to capture clusters with varying distributional shapes. We apply the clustering separately to even and odd pages, since there seem to be different generative principles at play, as evinced by the heat maps in Figure 4a & 4b. After estimating the optimal number of clusters for even and odd pages ($n = 8$ and $n = 7$), we fit a Gaussian Mixture model to the data.

The results of this clustering are presented in Figures 5a & 5b. These plots are based on a sample of 50,000 ads from even and odd pages. We see apparent differences in size and position in the newspaper between the odd and even pages. The odd pages contain more noise, especially for ads with larger sizes, whereas the ads on the even pages seems to be more structured. For some of these clusters, there is more variation, while others are closely clustered. A noteworthy cluster on the even pages is the brown cluster, which refers to full-page ads. We can also see that these ads appear toward the end of the newspaper. Newspapers are generally structured in columns, which is visible in the distinct structure present in the widths of advertisements (x-axis). The height of ads (y-axis) has more variance than the width, nevertheless, the algorithm distinguishes clusters based on their height. In terms of relative position, the number of advertisements

5 Scikit-Learn offers the Gaussian mixture algorithm for the estimation of a mixture model



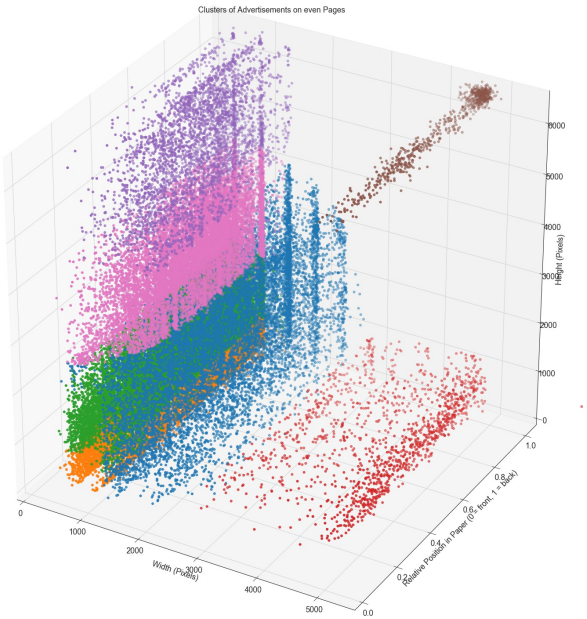
(a) Pixel Density of Advertisements on Even Pages in *Trouw*

(b) Pixel Density of Advertisements on Odd Pages in *Trouw*

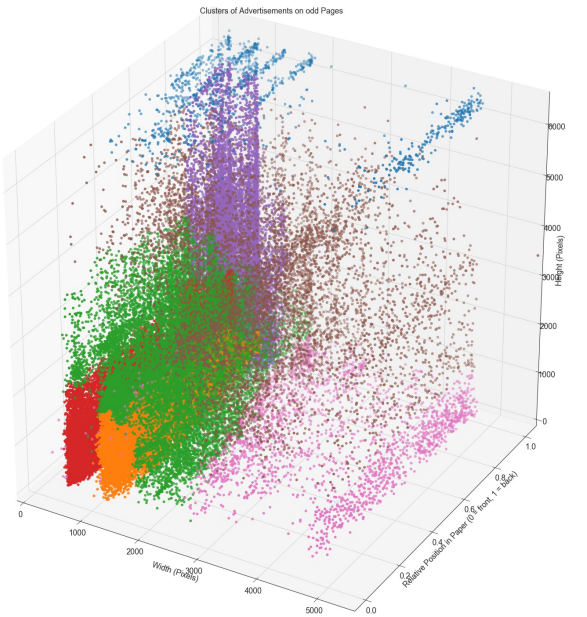
increases towards the back of the newspaper (z-axis). Ads with a small width and a more consderiable height, possibly classified ads, also appear toward the end of the paper.

The information contained in these clusters can help us filter out particular subsets of advertisements. For example, maybe we are not interested in classifieds ads or tiny ads. This method can help us determine the boundaries of these ads in terms of size and position.

We can also engineer additional features that help us refine our subsets even further. One useful feature, for instance, is character proportion, which can be calculcated by dividing the number of characters by the size (width x height) of the adverts. We can use this metric to filter out textual advertisements or those that predominantly consist of visual material. A similar feature is the proportion of digits, constructed by dividing the number of digits by the number of characters. This feature can be used to distinguish ads that mostly list prices of imported goods or wholesale goods, from other ads. Ostensibly, these numbers are dependent on the quality of the text extracted by Optical Character Recognition (OCR). Notwithstanding, variations between newspapers and differences over time, these numbers allow us to filter particular types of advertisements. For example, we can filter out ads by inspecting particular clusters in the distribution of the character proportion. This form of exploratory data analysis can, for example, help to quickly identify for a specific paper what advertisements consists of the columns of classifieds ads. These advertisements are characterized by a higher than average height and smaller width as well as a higher character proportion. We could also combine this metadata information with keyword search to further filter out types of ads for specific products.



(a) Clusters of Advertisement based on width, height and relative page number on even pages in *Trouw*



(b) Clusters of Advertisement based on width, height and relative page number on Odd pages in *Trouw*

In this section, we have demonstrated a possible approach to using metadata to learn more about the structure of advertisements in historical newspapers. There are ostensibly many more directions to explore. However, this section underscores that metadata should not be overlooked and can be used in conjunction with text analysis to guide further analysis.

3 Text Mining Advertisements

In this section, we show how we can extract textual information from historical advertisements, and how this textual information can be used for historical analysis of trends and particularities. More specifically, we present a case study on the advertised product nationalities of cigarettes. We construct our corpus of cigarette advertisement from a larger corpus of advertisements from ten national newspapers between 1890 and 1990 (Table 1).[6]

From this corpus of newspapers, we extracted a sub-corpus of advertisements that contain the three most common spelling variants of cigarette: *cigaret*, *sigaret*, and *cigarette*. These singular and plural variants are queried using the following regular expression: `'cigaret*\w+', 'sigaret*\w+'`. This query yielded 43,781 advertisements. Figure 6 displays the distribution of the relative number of these cigarette advertisements. The trend line shows that the relative number of cigarette advertisements grew until the early 1920s, after which it decreased. In the 1950s, the relative number of ads again peaked, and after the 1960s, the number of advertisements that included variants of the word cigarette dropped considerably, suggesting that after the 1960s, cigarette manufacturers advertised less in Dutch newspapers.

Plotting these time series gives an overview of the temporal distribution of advertisements for a particular product. However, the search terms should be carefully selected. Using the different spelling variations, captured with the regular expression, this query covers most of the historical variations in the twentieth century. In the case of cigarettes, the decrease in the last quarter of the twentieth century can in part be explained by the fact that advertisers stopped with explicitly referencing cigarettes. Rather, they turned to ads that include brand names or logos but not the word cigarette. In such instances, computer vision could help to detect cigarettes or the occurrence of particular logos in advertisements. Many of these brand names are ambiguous, and querying them could also return advertisements for unrelated products. Moreover, the period

6 Not all of these ten newspapers appeared throughout the entire period.

Newspaper	Period	Number of ads
Algemeen Handelsblad	1906-1970	979,312
Het Volk	1919-1945	191,626
Nieuwe Rotterdamsche Courant	1909-1929	472,536
NRC Handelsblad	1970-1990	460,996
Parool	1945-1990	1,626,204
Telegraaf	1893-1990	3,777,982
Trouw	1946-1990	1,154,746
Vaderland	1919-1945	317,440
Volkskrant	1940-1995	1,193,558
Vrije Volk	1945-1990	1,584,863

Tab. 1: Overview of selected national newspapers

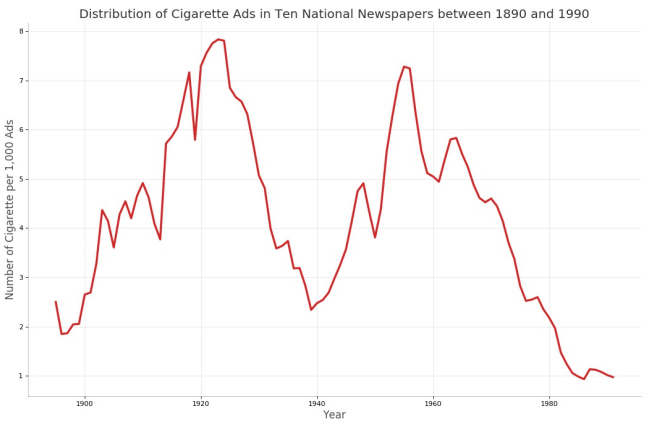


Fig. 6: Relative Number of Cigarette Advertisements 1890-1990

during the Second World War should also be scrutinized more closely, since there were only a few newspapers appearing throughout this period, and the size and content of newspapers transformed drastically. Even when using relative frequencies to offset the changing number of newspapers, we cannot account for such factors.

3.1 Cigarettes and Product Nationalities

While text mining gives us the ability to assess large-scale trends, it can also help to explore more particulate ways in which advertisers framed a product. One type of framing is the nationality of a consumer good, also known as product nationality. The relationship between products and constructions of nationality has long since been debated in consumer and marketing studies [Thakor, 1996, Hull, 2016]. A key find in these studies is that a strong connection to a favored country persuades consumers to spend more money on products with such a connection [Menapace et al., 2011]. The representations of product nationality are more than mere associations, they are "powerful narratives about the meanings and values transferred by products from their origin to their destination [Åskegaard and Ger, 1998]."

In the case of the cigarette, the representation of product nationalities is an under-researched topic, which is surprising given the fact that advertisers explicitly connected cigarettes to particular countries and regions. These product nationalities refer to their actual or perceived country of origin, but often they also communicated particular characteristics associated with these countries.

The advertisement in Figure 7 is a clear example of the association between product nationality and particular characteristics. This 1938 advert for the Buffalo cigarette linked the brand to the United States in several ways. First, the brand's name Buffalo denoted the emblematic American prairie animal as well as the city of Buffalo in upstate New York. The relationship between the brand and the United States was further enforced by a small print mentioning its producer: The Cumberland Company from Clarksville, USA. Second, in addition to these textual cues, the advertisement included a visual signifier: a background image of a giant cowboy bending over a Dutch tulip field. This picture of a cowboy—an exemplar of American culture—further substantiated Buffalo as an American cigarette. Furthermore, this image expressed the towering dominance of American products in the Netherlands. Third, the ad presented the Buffalo cigarette as having an American product nationality by describing it as "the tastiest and spiciest American cigarette." The geographical association to the

Fig. 7: Advertisements for Buffalo cigarettes in *Limburger Koerier*, April 13, 1938

United States suggested the product's country of origin, but also signified a particular taste specific to American techniques of tobacco preparation.

The Buffalo ad is of course just one example. How does this particular advertisement compare to others, and is its messaging specific to cigarettes with an American product nationality? Using text mining, we can gather the information that help us answer questions like these.

3.2 Charting Product Nationalities

We can, for example, chart the nationalities most commonly associated with cigarettes in advertisements. There are two basic ways to establish product nationality. First, we could count the occurrence of bigrams—two-word phrases—that include an explicit reference to nationality and cigarettes. For example, to find out when and how often advertisers described cigarettes as American, one could count the advertisements that contained bigrams such as American Cigarette (*Amerikaanse sigaret*). Regular expressions can help to capture a wide range of possible spelling variations of such a bigram. However, a key flaw in counting the bigram American cigarette is that we then only enumerate instances in which American appeared directly to the left of cigarette. Advertisers, however, also used other ways to relate cigarettes to a particular location. For instance, in the case of Egyptian cigarettes, advertisers relied on phrases such as "imported from Egypt." One possible yet time-consuming solution is to construct a list of all possible strings that express a relationship to a specific nationality.

A second method is to count references to nationality that co-occurred in the proximity of the word cigarette. In this approach, one only counts words that co-occur within a specific span of words, rather than advertisements that contain the two words. The co-occurrence of words in one advertisement does not necessarily indicate a relationship between them. American could appear in an advertisement for cigarettes without referring to the product nationality of the cigarette or one of its features. Nevertheless, proximity is a good indicator of a semantic relationship between words. Hence, in what follows, we count references to America within a span of five words to the right or left of the keyword cigarette.

One could also look for the co-occurrence of the two terms in one document without the use of a span. However, during digitization, the Optical Layout Recognition (OLR) did not always correctly compartmentalize advertisements. There are, for instance, cases where cigarette and America appeared in the actual newspapers in two separate advertisements, whereas after digitization, they were identified as one single advertisement (see Figure 8). In Figure 8, American refers


NAAMLOOZE VENNOOTSCHAP:

AMERICAN HOTEL,

gevestigd te AMSTERDAM.

Opgericht bij Akte, dd. 29 Maart 1882.

UITGIFTE EENER 5 PCTS. OBLIGATIE-LEENING,

groot ƒ 91,000,

aan te gaan krachtens machtiging der Buitengewone Algemeene Vergadering van Aandeelhouders, dd. 30 December 1882. (1701).

De INSCHRIJVING op deze 364 Obligatiën ad ƒ 250, à pari, met rente, ingaande 15 FEBRUARI 1883, is geopend op DONDERDAG den 18den JANUARI a. s., ten Kantore der Associatie-Cassa, van 's voormiddags 10 tot 's namiddags 4 uur.

PROSPECTUSSEN zijn te verkrijgen ten Kantore van de Heeren LÉON WERTHEIM & Cº., alhier.

Magazijn van Havana-Sigaren

C. BLOOKER Thzn.,

30, KALVERSTRAAT, 30. (1698).

OPRUIMING VAN DIVERSE SOORTEN SIGARETTEN.

Fig. 8: Example of an incorrect segmentation of advertisements. *Algemeen Handelsblad*, January 15, 1883

to the American hotel in Amsterdam, and *sigaretten* to an unassociated retailer advertising cigarettes. The words appeared in one single advertisement, albeit separated by a large number of words. Therefore, the use of a span of five words would not have counted this as an instance in which the two words appeared together. Looking for words in proximity to each other reduces the impact of errors produced by composite advertisements.

For this section, we charted six product nationalities associated with cigarettes: American, British, Egyptian, Russian, Turkish, and Virginia.[7] Using regular expressions, we queried the singular and plural variants of these references as well as common spelling variations.[8] The occurrence of these words in advertisements serves as a proxy for the popularity of cigarettes with varying product nationalities. Figure 9 shows the relative frequency of references to nationalities per 1,000 cigarette advertisements.

From 1890 to 1919, Egyptian, Turkish, and Russian cigarettes were the most popular. The popularity of these cigarettes mirrored the economic, cultural, and political power of the associated geopolitical entities. Before the First World War, the popularity of Russian and Turkish cigarettes mirrored the might of the Ottoman and Russian Empires. In the same period, the American cigarette industry was making its first forays into the European cigarette

7 We approach Virginia as a nationality, since the term came to represent Britain

8 Code is available on **XXX**

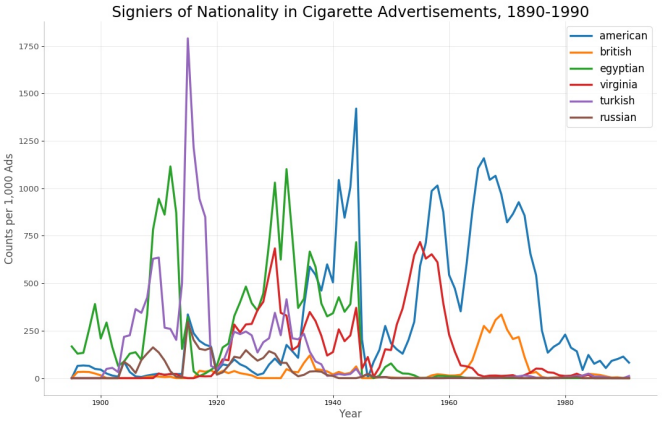


Fig. 9: Relative frequency of signifiers of nationality in cigarette advertisements, 1890-1990

market [Brandt, 2009]. Right after the First World War, Virginia and Egyptian cigarettes became the most popular cigarettes. After the First World War, the British cigarette industry, and especially British American Tobacco (BATCO), played a prominent role in the production and distribution of both Egyptian and Virginia cigarettes [Shechter, 2006, 27-8]. BATCO was also instrumental in disseminating Virginia cigarettes in the Netherlands.

The association with the United States gained prominence after the Second World War, when both American and Virginia cigarettes towered above the other nationalities, which by that time had disappeared almost entirely. In the 1960s, when Virginia cigarettes lost their popularity, the American cigarette acquired sole dominance. A decrease in references to the United States characterized the subsequent decades. This decline coincided with the growing sentiments of anti-Americanism in the Netherlands [Kroes, 2006, Gienow-Hecht, 2006]. Amid anti-American sentiments, advertisers might have refrained from associating their product with the United States. Furthermore, in 1964, the American Surgeon General Luther Terry published the Report on Smoking and Health in which he presented the detrimental effects smoking could have on one's health. The report led to a significant decrease in cigarette consumption in the Netherlands and the United States [Hoffmann et al., 2001].

As this section has shown, counting specific strings of text in digitized material is a relatively easy and fast way to gauge and compare the popularity of particular products. The trends in popularity in Dutch newspapers matched

global trends. However, the interest in American cigarettes already waxed before the Second World War, while academic literature on Americanization in the Netherlands commonly situates this process after the Second World War.

A next step could be to examine how particular products were presented in advertisements. In other words, were there distinct product features for cigarettes with particular product nationalities? Understanding the historical use product features in advertising discourse can help us determine the cultural and technological impact of products.

3.3 Product features

In this section, we discuss two methods to examine the historical evolution of particular characteristics associated with cigarettes. First, we show how an algorithm to detect bursty word use can be leveraged to extract features that typified a certain period. Second, we demonstrate how machine learning can determine whether cigarettes with a clear product nationality possessed features distinct from cigarettes with other nationalities.

3.3.1 Finding Trending Topics using Burst Detection

When we want to study the use of particular words and examine how this use has evolved, one of the first tasks is to decide which words we are actually studying? One method is to rely on secondary sources to construct a vocabulary of words related to a subject or particular period. A different approach is a data-driven one, in which we extract words that exhibit noteworthy use in the corpus using algorithms. One such algorithm is burst detection.

Burst detection is a modeling technique to detect *bursts of activity* in streams of data, for example, the sudden rise and fall in word frequency in serial publications. The trends for individual words cannot be just be compared in a time series. For example, for words with relatively little overall activity, a sudden, repetitive increase in use can signal a burst, whereas, for words with much activity, a different burst intensity might be required. Moreover, bursts of activity can also be nested within larger patterns. To be able to capture these bursts, Jon Kleinberg developed an algorithm that models the stream of information as an infinite-state automaton. The algorithm assigns costs to state transitions, which makes it possible to distinguish between short bursts and long burst even while the overall rate of transmission changes over time [Kleinberg, 2002]. We use Kleinberg's algorithm to detect whether and when specific words exhibited

"bursty" behavior in advertising discourse. In other words, we use it to identify *trending topics* in advertising discourse.

We model the burstiness for a subset of words. This subset includes 500 most distinctive, determined using tf-idf, adjectives and nouns from the corpus of cigarette advertisements. We then apply Kleinberg's algorithm using the default settings on monthly frequency counts of these 500 words. For each of these words, we get information on whether they bursted and how intense the burst was, as well as the duration of the burst. Figure 10 displays the top 50 bursty nouns and adjectives and when they bursted. We can see the appearance of the keyterm filter around 1965 when the Surgeon General report also appeared.

Moreover, the debate in the late 60s and early 1970s shifted toward the amount of nicotine in cigarettes. The appearance of the term health (*gezondheid* in the mid 1920s stems from advertisements that promoted the health benefits of smoking Virginia cigarettes. The figure also contains phrases such as job application (*sollicitatie* and human resources (*personeelszaken*). Upon closer inspection, these words appeared in job ads placed by cigarette companies, a category that researchers might want to prune from their corpus.

Using this method, we can quickly gauge when particular words displayed "bursty' behavior or which topics were trending at particular moments in time. This information can subsequently be used for closer examination of the texts in which these words appeared. Techniques like this can also help to distinguish a particular subset of that one might want to remove or treat as a separate corpus.

3.3.2 Detecting Distinctive Features Using Text Classifiers

A different line of questioning might focus on the differences in vocabulary between cigarettes with varying product nationalities. For example, were American cigarettes more commonly associated with a mild taste, and were Egyptian cigarettes presented as artisanal? One method to gauge the differences between corpora is through the use of a machine learning classifier. These classifiers learn which textual features are predictive of a certain class. For example, which words are indicative of American cigarettes and which for Virginia cigarettes. This method assumes that language use between the two stays as separated over time. In other words, if features are first introduced in Virginia cigarettes and then co-opted by American cigarettes, it might be more challenging to use this feature to distinguish between the two. For this reason, we can also train classifiers for separate periods, given that we have enough data. If the classifiers fail to separate the corpora, we can infer that the differences between language use were not very

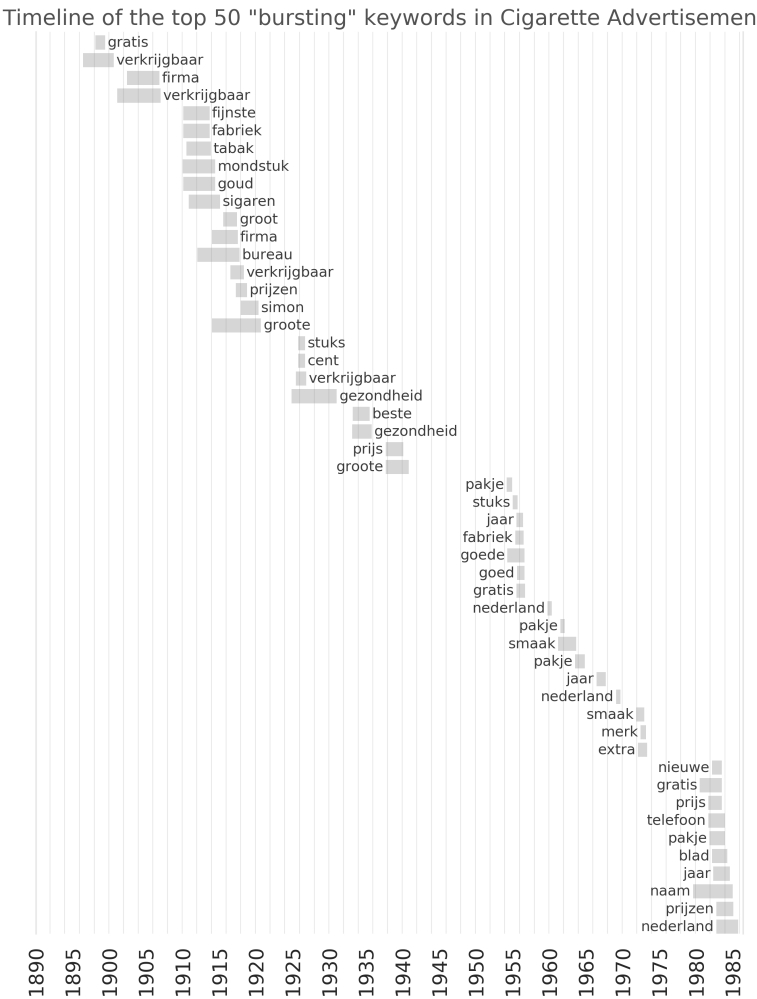


Fig. 10: Top 50 Bursty Nouns and Adjectives in Cigarette advertisements, 1890-1990

large. For each of these periods, we can also investigate the most informative features.

For this method, We divide the corpus into subcorpora of ads that contain either references to the United States or to Virginia and the United Kingdom. After removing explicit references to nationalities, we train a Naive Bayes Classifier on the two corpora. For the period 1920-1980, in which both types of cigarettes were represented, the classifier is able to distinguish between advertisements for American and Virginia cigarettes with an accuracy of: 0.88, and respective F_1-scores of .92 and 0.79.

Subsequently, we can examine the most informative features for the classifier to label an ad as either American or Virginia. Noteworthy results include brand names such as Lucky Strike, Roxy and Camel for American cigarettes, and Derby and Chief Whip for Virginia cigarettes. Moreover, words such as cork (*kurk*), mouth piece (*mondstuk*), purity (*zuiverheid*), and health (*gezondheid*) were predictive for Virginia cigarettes, while connaisseurs (*liefhebbers*), packages (*pakjes*), filter, and smoking pleasure (*rookgenot*) are all predictive of American cigarettes.

The Ardath Company was primarily responsible for connecting these aspects to Virginia cigarettes; the British cigarette manufacturer boasted that the purity of its Virginia tobacco led to a better tasting and healthier cigarette. This link was particularly strong in advertisements for the brand Chief Whip, which Ardath described as the "zenith of purity. Together with the company Wills, Ardath distinguished the taste of Virginia cigarettes from that of American cigarettes. Ardath and Wills both denounced saucing and blending——two key features of the American cigarettes——and distanced themselves from American cigarettes in doing so. In 1925, Wills claimed that its Virginia cigarettes consisted of 100% pure Virginia tobacco, without the addition of Greek, Indonesian, or Turkish tobaccos. In advertisements for Chief Whip, a doctor claimed that the cigarette was "absolutely pure and free of all surrogates and sauces." Advertisers presented Chief Whip as a pure and unprocessed cigarette.

The features used by the classifier point to three distinctive material aspects of the American cigarette: its length, its filter tip, and its packaging. Two addition that American cigarette producers introduced after the Second World War. First, in the 1950s, the longer, king-size cigarette was introduced to Dutch consumers, after enjoying great success in the United States. Advertisers linked the increase in length to the United States to help familiarize the Dutch smoker with the longer cigarette. In 1955, So Long referred to the United States in their explanation of what a king size cigarette was: "In America, a cigarette longer than 85mm

is called King Size."[9] Advertisers used the link with the United States to help acquaint Dutch consumers with the longer cigarettes.

The second significant change was the introduction of filter cigarettes. Amid growing health concerns in the United States, American cigarette manufacturers introduced the purportedly healthier filter cigarettes in the early fifties [Brandt, 2009, 244]. In the context of American filter cigarettes, purity did not denote the unblended nature of cigarettes, but the purifying effects of filters. Advertisements claimed that a lengthy filter would lead to "more and purer smoking."[10] Filter cigarettes became hugely popular, and by the 1970s, almost ninety percent of the cigarette market consisted of filter cigarettes [Brandt, 2009, 244].

Third, technological developments in the United States changed the look and feel of cigarette packaging. These innovations offered advertisers new ways to link the product to the United States. American companies were the first to package cigarettes mechanically in plastic-wrapped cardboard boxes. Until then, consumers bought cigarettes per piece and stored them in less practical tin boxes. The new method of packaging not only referred to the United States because of its origin, but it also carried cultural connotations that resonated with American culture. Advertisers described the packaging of the American cigarette as flat, practical, modern, famous, or fancy.

The information gained through this method can be used for closer examinations of particular characteristics, as demonstrated above. A more in-depth analysis falls outside of the scope of this chapter.[11]

4 Conclusion

Advertisements are a rich and varied source for the study of public discourse. This chapter showcases how we can apply computational methods to metadata information to learn more about advertising in the past. The ability to distinguish between different types of advertisements can help to streamline further inquiry. Moreover, based on a case study on product nationalities in cigarette advertisements, we showed how text mining can help to better understand the historical trajectories between products and particular nationalities. Also, we demonstrated how the analysis of bursty words and machine learning can help

9 "So Long advertisement," *Het Vrije Volk*, December 13, 1955.
10 "Sir Richard advertisement," *De Telegraaf*, January 5, 1962.
11 see [Wevers, 2017] for such an analysis

to identify trending topics and distinctive words in advertising discourse. While the text is often far from perfect in digitized newspapers, there is still much we can extract from this abudance of source material using computational methods. With the increasing amount of digitized newspapers collections in different countries, we can also expand our efforts to study advertisements in a transnational perspective.

Acknowledgment: The author acknowledges the National Library of the Netherlands (KB) for making its newspaper data available.

References

[Arnold and Tilton, 2019] Arnold, T. and Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities*.

[Åskegaard and Ger, 1998] Åskegaard, S. and Ger, G. (1998). Product-country images: Towards a contextualized approach. *European advances in consumer research*, 3(1):50–58.

[Bilgin et al., 2018] Bilgin, A., Hollink, L., van Ossenbruggen, J., Sang, E. T. K., Smeenk, K., Harbers, F., and Broersma, M. (2018). Utilizing a Transparency-driven Environment toward Trusted Automatic Genre Classification: A Case Study in Journalism History. *arXiv:1810.00968 [cs]*.

[Brandt, 2009] Brandt, A. (2009). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product That Defined America*. Basic Books, New York.

[Chung et al., 2015] Chung, J. S., Arandjelović, R., Bergel, G., Franklin, A., and Zisserman, A. (2015). Re-presentations of art collections. In Agapito, L., Bronstein, M. M., and Rother, C., editors, *Computer Vision - ECCV 2014 Workshops*, pages 85–100, Cham. Springer International Publishing.

[Daems et al., 2019] Daems, J., D'haeninck, T., Hengchen, S., Zere, T., and Verbruggen, C. (2019). 'Workers of the World'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940. *Journal of European Periodical Studies*, 4(1):99–114.

[Fox, 1997] Fox, S. R. (1997). *The Mirror Makers: A History of American Advertising and Its Creators*. University of Illinois Press, Urbana.

[Gienow-Hecht, 2006] Gienow-Hecht, J. C. E. (2006). Always Blame the Americans: Anti-Americanism in Europe in the Twentieth Century. *The American Historical Review*, 111(4):1067–1091.

[Hoffmann et al., 2001] Hoffmann, D., Hoffmann, I., and El-Bayoumy, K. (2001).

The less harmful cigarette: A controversial issue. A tribute to Ernst L. Wynder. *Chemical research in toxicology*, 14(7):767–790.

[Hull, 2016] Hull, G. (2016). Cultural Branding, Geographic Source Indicators and Commodification. *Theory, Culture & Society*, 33(2):125–145.

[Kleinberg, 2002] Kleinberg, J. (2002). Bursty and Hierarchical Structure in Streams $_*$. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 25.

[Kroes, 2006] Kroes, R. (2006). European Anti-Americanism: What's New? *Journal of American History*, 93(2):417–432.

[Lears, 1994] Lears, T. J. (cop. 1994). *Fables of Abundance: A Cultural History of Advertising in America*. Basic Books, New York.

[Marchand, 1985] Marchand, R. (cop. 1985). *Advertising the American Dream: Making Way for Modernity, 1920-1940*. University of California Press, Berkeley.

[Marshall, 1995] Marshall, M. (1995). *Contesting Cultural Rhetorics: Public Discourse and Education, 1890-1900*. University of Michigan Press, Ann Arbor.

[Menapace et al., 2011] Menapace, L., Colson, G., Grebitus, C., and Facendola, M. (2011). Consumers' preferences for geographical origin labels: Evidence from the Canadian olive oil market. *European Review of Agricultural Economics*, 38(2):193–212.

[Postman, 2005] Postman, N. (2005). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Penguin, London.

[Schudson, 1982] Schudson, M. (1982). *The Power of News*. Harvard University Press, Cambridge.

[Shechter, 2006] Shechter, R. (cop. 2006). *Smoking, Culture and Economy in the Middle East: The Egyptian Tobacco Market 1850-2000*. Tauris, London.

[Thakor, 1996] Thakor, M. (1996). Brand origin: Conceptualization and review. *Journal of Consumer Marketing*, 13(3):27–42.

[van Eijnatten and Ros, 2019] van Eijnatten, J. and Ros, R. (2019). The Eurocentric Fallacy : A Digital-Historical Approach to the Concepts of 'Modernity', 'Civilization' and 'Europe' (1840–1990). *International Journal for History, Culture and Modernity*, 7:686–736. Accepted: 2020-02-04T17:16:10Z Library Catalog: dspace.library.uu.nl Pages: 686-736.

[van Strien et al., 2020] van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks:. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

[van Vree, 1989] van Vree, F. (1989). *De Nederlandse pers en Duitsland 1930-*

1939: een studie over de vorming van de publieke opinie. Historische Uitgeverij, Groningen.

[Wevers, 2017] Wevers, M. (2017). *Consuming America: The United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890-1990*. Ph.D. dissertation, Utrecht University, Utrecht.

[Wevers and Smits, 2020] Wevers, M. and Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1):194–207.