

RADBOUD UNIVERSITY NIJMEGEN



FACULTY OF SOCIAL SCIENCES

Periods on Paper

TOPIC MODELING AND SENTIMENT ANALYSIS OF MENSTRUATION IN 20TH-CENTURY DUTCH
NEWSPAPERS

THESIS MSc ARTIFICIAL INTELLIGENCE

Author:
Anna LEBBINK
s1105054

Supervisor:
Prof. Martha LARSON

Second reader:
Dr. Iris HENDRICKX

August 2025

Abstract

Menstruation remains a stigmatized topic in many societies. Newspapers shape and reflect public opinion, yet there is little research on the coverage of menstruation in Dutch newspapers. This thesis addresses that gap by using topic modeling and sentiment analysis to explore how menstruation was discussed in the Netherlands throughout the twentieth century. Using menstrual-related keywords, 9712 relevant articles were retrieved from the Delpher newspaper archive, and the influence of OCR errors and the disambiguation of keywords was assessed. With Latent Dirichlet Allocation (LDA), seventeen distinct topics were identified, including abortion, contraceptives, menopause, religion, and women's health. The sentiment analysis, using a fine-tuned version of RobBERT, revealed 7419 neutral, 1939 positive, and 354 negative newspaper articles. The ideological background (pillarization) of newspapers influenced how menstruation was covered, both in terms of topic and sentiment. However, whether a newspaper was national or local had minimal impact. This research highlights how computational methods can uncover historical patterns in sensitive or stigmatized topics.

Contents

1	Introduction	4
2	Background	6
2.1	Menstruation and Reproductive Health	6
2.2	The History of Menstrual Taboos	6
2.2.1	Global Taboos Around Menstruation	7
2.2.2	Menstrual Taboos in the Netherlands	7
2.3	Newspapers	8
2.3.1	Delpher: The Dutch National Library	9
2.3.2	Verzuiling (Pillarization)	9
2.3.3	World War II and Press Censorship	10
2.4	Menstruation in Media Discourse	10
2.4.1	Social Media	10
2.4.2	Newspapers Worldwide	10
2.4.3	Menstruation in Dutch Media	11
3	Related Works	11
3.1	Optical Character Recognition	12
3.2	Word Sense Disambiguation	12
3.3	Text Mining	12
3.3.1	Topic Modeling	12
3.3.2	Sentiment Analysis	13
4	Data	14
4.1	Data Design	14
4.2	Data Acquisition	15
4.2.1	Keyword Selection	15
4.2.2	Ambiguous Keywords and Contextual Challenges	15
4.2.3	Article Collection and Filtering	17
4.3	Data Preprocessing	18
4.3.1	Word Sense Disambiguation	19
5	Topic Modeling	20
5.1	Method	20
5.2	Results and Discussion	21
5.3	Discussion	25
6	Sentiment Analysis	26
6.1	Method	26
6.2	Results & Discussion	27

6.2.1	Discussion	29
7	Pillarization & Geographical Scope	29
7.1	Method	31
7.2	Results	31
7.2.1	Scope and Pillars	32
7.2.2	Topic Modeling	33
7.2.3	Sentiment Analysis	37
7.3	Discussion	38
8	Discussion	39
8.1	Interpretations	40
8.2	Limitations	42
8.3	Recommendations	42
9	Conclusion	43
10	Acknowledgments	43
11	Appendix	48
A	Topic Modeling Results	48
B	Newspaper Pillars & Scope	50
B.1	Frequency Pillars per Topic	50
B.2	Newspapers Metadata	51

1 Introduction

Menstruation is experienced by half the population for a substantial part of their lives. Yet, across many cultures and histories, it is enveloped in silence, stigma, and misunderstanding. In the Netherlands, too, menstruation has often been treated as a private matter, rarely discussed openly in public.

For example, an anecdote from my grandmother: She and my grandfather had a drugstore in The Hague in the 1970s and 1980s, she told me, that when people, and mostly only women, came to buy sanitary pads, they were always put in a separate opaque bag, so it could be hidden. This lack of visibility contributes to persistent myths, silence, and social exclusion.

This thesis investigates how menstruation was discussed in Dutch newspapers during the twentieth century. Newspapers are an important part of cultural history, which shape and reflect societal values. Before the widespread adoption of radio, it took until the 1960s for every household to have one, and the television, until 1970, newspapers were among the most influential forms of mass communication. They offer a valuable lens for studying the shifts in public attitudes and discourse.

While previous studies have explored the portrayal of menstruation in modern media, advertising, and education [16, 20, 4, 42], there is a gap in the literature when it comes to how menstruation was historically discussed in Dutch newspapers. No study has yet explored this topic, particularly using large-scale digital analysis. This thesis aims to fill that gap by analyzing Dutch newspaper articles by focusing on evolving themes, sentiments, ideological background, and geographical scope related to the menstrual cycle.

The corpus for this research is gathered from Delpher, the digital archive of the Dutch National Library, which contains millions of digitized historical newspapers, magazines, and books. These newspapers are scanned and processed via Optical Character Recognition (OCR). However, the OCR quality varies, particularly for older or damaged documents, affecting the accuracy and completeness of data extraction. Preprocessing techniques are used to clean and improve the dataset before analysis.

Given the evolving language around menstruation, the selection of keywords required extra consideration. For example, the Dutch term *ongesteld*, now a commonly used term for menstruation, was historically used more broadly to mean ‘unwell.’ To address such ambiguities, word sense disambiguation techniques were employed, like text classification models trained to distinguish menstrual contexts from non-related uses.

This study applies Natural Language Processing (NLP) techniques to explore two key aspects of the data: topic modeling and sentiment analysis. Using Latent Dirichlet Allocation (LDA), the goal is to identify and trace the dominant themes associated with menstruation over time. In parallel, sentiment analysis is conducted using RobBERTa, a transformer-based language model fine-tuned for Dutch sentiment classification. This enables analysis of how emotional or sentimental language surrounding menstruation evolved across decades.

A core aspect of this study is ideological. Dutch society in the twentieth century was characterized by *verzuiling*, or pillarization, a system that divided institutions, including newspapers, along ideological lines: Catholic, Protestant, Socialist, and Liberal. Each pillar maintained its own schools, political

parties, and media outlets. By comparing how different ideological and geographical perspectives covered menstruation, this study explores how political and religious beliefs shaped public discourse on female reproductive health.

Together, the combination of topic analysis, sentiment evaluation, ideological, and geographical comparison can paint a picture of how menstruation was covered in Dutch newspapers throughout the twentieth century.

Understanding how menstruation was historically talked about is important because the media plays an essential role in shaping and reflecting public opinion. Topics that were widely discussed, or deliberately avoided, show society's attitude and the extent of the stigma over time. By comparing newspapers with different ideological and geographical scopes, this research aims to discover how religion, politics, and regional differences influenced the reporting of the menstrual cycle in Dutch newspapers during the twentieth century.

Research Question:

What topics and sentiments about the menstrual cycle were expressed in Dutch newspapers in the twentieth century?

Sub questions:

1. To what extent does OCR quality affect the accurate retrieval of menstruation-related newspaper articles from the Delpher archive?
2. How can semantic ambiguity in menstruation-related keywords (e.g., *ongesteld*, *maandstonde*, *maandverband*, *tampon*) be addressed in historical newspaper analysis?
3. What topics related to the menstrual cycle were discussed in Dutch newspapers during the twentieth century, and how have these topics changed over the years?
4. What sentiments did Dutch newspapers express on each topic, and how did these sentiments change over time?
5. How do topics and sentiments on the menstrual cycle vary across newspapers with different ideological backgrounds (Protestant, Roman Catholic, Socialist, Liberal) and geographical scopes (local/national)?

By applying NLP methods to historical data, this study shows how digital techniques can uncover large-scale patterns in newspapers. It contributes to multiple academic fields, including digital humanities, natural language processing, media history, and gender studies. In particular, it demonstrates how topic modeling and sentiment analysis can be applied to low-resource, OCR-affected historical texts, particularly when examining stigmatized or under-researched topics.

Moreover, this thesis provides insight into ideological bias and social silence in historical media. By analyzing developments such as the introduction of menstrual products and the contraceptive pill, it highlights how technological and political shifts influenced public narratives.

2 Background

This chapter provides the background on the social, cultural, and medical context in which menstruation has been discussed. It first introduces the biological aspects of the menstrual cycle and related health conditions, followed by an exploration of menstrual taboos both globally and within the Netherlands. It then turns to the historical role of Dutch newspapers in shaping public discourse and concludes with a discussion of the relevance and limitations of the Delpher archive.

2.1 Menstruation and Reproductive Health

The menstrual cycle is a biological process experienced by individuals with a uterus, encompassing almost half the global population. The first menstruation, also called menarche, happens between 12 and 15 years of age, and the last menstruation is around 50 years of age, marking the start of menopause. Contrary to common belief, menstruation is only one phase of a continuous hormonal cycle.

A typical menstrual cycle spans approximately 28 days, though cycles ranging from 21 to 35 days are also considered normal. The cycle is divided into four phases: menstruation (menses), follicular, ovulation, and luteal. During the menses phase, lasting 3 to 7 days, the uterine lining is shed. The follicular phase, which overlaps with menstruation, is marked by rising estrogen levels and follicle development, leading to ovulation around day 14. Ovulation involves the release of a mature egg, and the luteal phase follows, preparing the body for a potential pregnancy through increased progesterone levels. If fertilization does not occur, hormone levels drop, and the cycle restarts.

Physical and psychological symptoms such as cramping, bloating, acne, and mood swings are common. However, extreme symptoms may indicate underlying conditions such as endometriosis, polycystic ovary syndrome (PCOS), premenstrual dysphoric disorder (PMDD), or postural orthostatic tachycardia syndrome (POTS). These conditions are often underdiagnosed due to widespread stigma and a historical lack of research into women's health.

A major contributing factor to this underdiagnosis is the underrepresentation of women in medical and clinical studies, which has led to critical knowledge gaps in understanding female health, including how hormonal cycles influence drug efficacy and safety [9]. Drug effects can vary between the sexes due to differences in body composition, size, and pharmacological responses. This can lead to inappropriate dosing, inaccurate estimation of side effects, and ultimately less effective treatment for women. The lack of research on menstrual-related health conditions means that many treatments are not optimized for women, leading to poorer clinical outcomes and overall lower quality of healthcare.

2.2 The History of Menstrual Taboos

A taboo is a social or cultural rule that forbids discussing certain topics or engaging in certain behaviors that are often seen as offensive or inappropriate. These norms are reinforced through societal expectations and institutions such as religion, education, and media. While this thesis mainly examines how menstruation was depicted in Dutch newspapers, it is important to place these findings within a broader global context. The next two sections explore taboos around the world and in the Netherlands.

2.2.1 Global Taboos Around Menstruation

In many cultures and religions, menstruation is considered taboo, with many surrounding stigmas that result in restrictions on daily activities, participation in rituals, or social interactions. These restrictions are perpetuated by inadequate menstrual education [7] and strong cultural beliefs.

Two religions that hold such negative views toward menstruation are Hinduism and Orthodox Judaism [8, 21]. During menstruation, individuals may be prohibited from touching food, partaking in rituals, or engaging in sexual intercourse. In some Hindu traditions, men are told not to have conversations with a woman during her menstruation. While more modern and liberal denominations of these religions may have more lenient views towards menstruation, it still has a large influence on the daily lives and experiences of many menstruating individuals.

Education plays a significant role in how society looks at the menstrual cycle. If menstruation is addressed in schools at all, it is still often incomplete, inaccurate, or taught in a gender-segregated manner [5]. When formal education is absent, the information typically comes from family members, television, or online platforms, but this is also often incorrect and incomplete [19]. In some cultures, such as in India, conversations around menstruation are often avoided, making it difficult for individuals to seek support or access appropriate healthcare resources.

The widespread stigma and taboo around menstruation negatively impact individuals in multiple aspects of life. The lack of access to menstrual products [32], fear of embarrassment [39], or severe menstrual pain (dysmenorrhea) [34] can force girls to stay home from school, or in some extreme cases, drop out entirely [33]. This issue extends into adulthood, where dysmenorrhea and symptoms cause individuals to take sick leave from work [28].

Later in life, the start of menopause introduces more challenges. Many women report feeling vulnerable due to the stigma surrounding this transition [29]. Due to misinformation and stigma, schools and workplaces often fail to accommodate these needs, leading to concealment and distress among individuals experiencing the menstrual cycle.

2.2.2 Menstrual Taboos in the Netherlands

The presence of the menstrual taboo in the Netherlands is evident in both language and education. Vancauwenbergh & Franco (2023) [38] analyzed common Dutch expressions related to menstruation and found that many are either humorous or derogatory, which reinforces discomfort and trivialization of the topic in everyday language. Zuidema et al. (2025) [43] explored how Dutch women perceive menstrual education and concluded it to be insufficient and outdated. The study emphasizes that inadequate education reinforces societal stigma and calls for a reevaluation of menstrual education by both schools and parents.

Lonkhuijzen et al. (2023) [37] did a study on how the people of the Netherlands look at menstruation and sexual acts during menstruation. They found that stigmatized attitudes resulted in constant physical and mental menstrual management, negatively impacting the lives of menstruating individuals. They conclude that improving menstrual education could have a destigmatizing effect.

Zedelijkheidswet In 1911, the Dutch government enacted the *Zedelijkheidswet* (Morality Act), which legally prohibited various forms of what was considered indecent behavior. The act banned the distribution of pornography, prostitution, contraception, and the public sale of sex toys. Sexuality became a taboo subject; there was hardly any talk in public, even by medical professionals. Particularly within the Roman Catholic Church, there was a fear that open discussion of sexuality and related subjects would lead to too much debauchery. As a result, sexuality was treated as a private matter, and menstruation and reproductive health were seldom publicly discussed.

The Pill and Abortion The contraceptive pill was introduced in the Netherlands in 1963, but it was initially only prescribed to married women as a means of regulating menstrual cycles instead of birth control. It was not until 1969, after the *Zedelijkheidswet* was abolished, that the pill could be described as a contraceptive pill. Despite this legal change, access remained limited, as prescriptions were often denied by doctors due to personal or religious beliefs [44]. The restricted access to contraceptives and the silence surrounding menstruation show the broader cultural and religious attitudes of the time and the discouragement of open discussion about female reproductive health.

Abortion had been illegal in the Netherlands since before the Morality Act, but in 1911 the law made it such that the penalty for illegal abortion was up to three year imprisonment. In the 1960s and 1970s, the rise of the women’s emancipation movement brought reproductive rights into the spotlight. Slogans such as “*Baas in eigen buik*” (“Boss of my own belly”) and “*De vrouw beslist*” (“The woman decides”) become prominent during protests. After years of debate, the *Wet afbreking zwangerschap* (Termination of Pregnancy Act) was passed by a narrow margin and went into effect on November 1, 1984.

Nederlandse Vereniging voor Seksuele Hervorming (NVSH) The Nederlandse Vereniging voor Seksuele Hervorming (Dutch Society for Sexual Reform), or NVSH, is a Dutch sexual advocacy organization [26]. It started as Neo-Malthusian League in 1881 as the first birth control clinic in the Netherlands. In 1946, the organization changed its name and broadened its goals. The NVSH has long advocated for gender equality, accessible contraception and abortion services, and comprehensive sexual education.

2.3 Newspapers

Today, discussions about menstruation often occur on social media, which has become a dominant platform for sharing information and personal experiences. However, newspapers remain an important way of sharing information, providing news, research findings, and opinion pieces. In the Netherlands, radio and television did not become common for every household until the 1960s and 1970s, respectively. Social media became widely used only in the early 2000s. As a result, newspapers were the primary source of information in the first half of the twentieth century and remained highly influential in the second half. The following subsections cover the archival and historical significance of newspapers in the Netherlands.

2.3.1 Delpher: The Dutch National Library

This thesis uses digitized newspaper articles from *Delpher*, the digital archive of the Dutch National Library. Delpher is an extensive online repository that offers access to historical Dutch newspapers, magazines, and books dating from the sixteenth century to the present. Many of the newspapers are scans of the physical copy and processed using Optical Character Recognition (OCR). Delpher works together with various external archives to digitize these materials. However, there are some gaps in Delpher's database, which could be due to several reasons. Some gaps may be due to a newspaper issue never having been saved, or the state of the historical newspapers. In other cases, it could be that they have not yet been turned into digital data, or are only available in other archives; however, Delpher is continuously working to rectify this.

2.3.2 Verzuiling (Pillarization)

An important feature of Dutch society in the twentieth century was *verzuiling*, or pillarization. This system divided institutions into distinct ideological 'pillars': Roman Catholic, Protestant, Liberal, and Social Democratic. Each pillar had its own political parties, schools, broadcasting systems, and newspapers. As a result, newspapers catered to their respective ideological communities and often reflected their values, beliefs, and social norms.

This ideological segmentation makes it possible to analyze how different groups represented menstruation in public discourse, in ways that aligned with their worldview. While these four were the major pillars, not all newspapers fell neatly into one of them. Some newspapers were small, local publications that did not report from a clear ideological perspective. Others represented alternative ideological positions such as Communism, Judaism, or politically neutral standpoints.

The way menstruation was discussed likely varied across different newspapers, depending on their ideological background and geographical scope. For example, a Catholic newspaper might present a different perspective on menstruation compared to a Socialist or Liberal newspaper, influencing both the topics covered and the sentiment expressed.

Examples of local newspapers are *Nieuwsblad van het Noorden* (Socialist) and *Het Vaderland* (Liberal). Examples of national newspapers are *De Volkskrant* (Roman Catholic), *Trouw* (Reformed Protestant), *Het Parool* (Socialist), and *Algemeen Dagblad* (Liberal).

While the influence of pillarization began to decline in the mid-1960s, partly due to secularization and the diminishing authority of the church, its effects remained in Dutch society and media. This ideological shift also influenced the newspapers themselves, as some gradually changed or abandoned their original stance. However, for this study, these ideological shifts are not taken into account. Instead, each newspaper is classified according to its original pillar affiliation, for the sake of methodological clarity and consistency.

2.3.3 World War II and Press Censorship

During the German occupation (1940-1945), press freedom in the Netherlands was suppressed. Newspapers were censored or shut down, and underground publications emerged to provide independent information. After the war, collaborating newspapers were banned or penalized. Delpher includes many newspapers from this period [1].

This censorship likely affected the amount and type of menstrual-related news articles during the war years. Due to editorial restrictions, sensitive topics such as reproductive health, women's rights, and bodily autonomy may have been intentionally avoided, downplayed, or presented in line with occupying ideology. As a result, the decrease or change in topic frequency and sentiment during this time may reflect not only societal shifts but also the lack of journalistic freedom.

2.4 Menstruation in Media Discourse

Media plays a central role in shaping how menstruation is discussed in public conversations. From traditional newspapers to social media, each platform presents different narratives about menstruation. This section explores how menstruation is currently represented in media and how it historically appeared in both international and Dutch newspapers.

2.4.1 Social Media

In current times, social media platforms have enabled the open discussion of menstruation, which creates a feeling of anonymity. Individuals, often young people, use platforms such as Twitter (X) or Reddit to seek advice, share experiences, and discuss topics such as menstrual hygiene products, cycle irregularities, and pain management.

However, the discourse is not always positive. A study by Davies et al. (2022) [10] analyzed Twitter discussions and found that there was an overwhelming emphasis placed on the negative expectations and shame around menstruation. Only a minority of tweets focused on advocacy or education. This suggests that while social media has opened up new avenues for dialogue, the discussions remain predominantly negative.

Similarly, Tomlinson (2021) [35] examined how internet memes are used in menstrual discourse. The study concluded that despite their humor, memes often reinforce negative stereotypes and reflect the continuous normalization of menstrual discomfort and embarrassment in everyday life.

2.4.2 Newspapers Worldwide

Print and digital newspapers have traditionally played a significant role in shaping public opinion. However, the limited research on mentions of menstruation in newspapers and the findings of existing studies suggest that menstrual topics are ignored and under-reported. This section reviews these works.

One study found that menstrual-related information was lacking in mainstream media, particularly the long-term effects of oral contraceptive use [20]. The study also discovered that this absence of information was neither addressed nor challenged by health professionals. This points to the broader pattern

of neglect in representing women’s health issues, potentially stemming from the underrepresentation of women in clinical studies.

When menstrual-related subjects are addressed in newspapers, it is often part of a discussion or “taboo-breaking”. A master thesis analyzed the discourse of the tampon tax in French media [16]. The study observed that menstruation was discussed in either economic or feminist frames. While these seem different, the study found that both reproduce gender stereotypes and depict menstrual blood as a social threat. The thesis concludes that the media cannot challenge the taboo of menstruation as long as it represents menstruation as a threat to society.

There are two specific moments when menstruation was (indirectly) mentioned in newspapers. In 2015 and another in 2020, two tennis players attributed their losses to “girls’ things”, avoiding explicitly mentioning it. A study analyzed news articles from around these times and found that in 2015, discussing menstruation in the media was described as both brave and shocking, with frequent mentions of the word taboo. By 2020, such discourse largely vanished, but the reluctance to name menstruation explicitly persisted. This suggests that while media framing may evolve, the underlying culture of concealment remains.

2.4.3 Menstruation in Dutch Media

Despite increasing awareness, menstruation remains a controversial topic in Dutch media. The 2017 #bloodnormal advertising campaign by the sanitary pad brand Libresse, which used red liquid instead of blue to depict menstrual blood, was considered “taboo-breaking.” Nieuwendijk (2019) [25] analyzed the reactions of the audience to the #bloodnormal campaign and found that the viewers assigned different meanings to it on three levels: the functional, commercial, and cultural. The campaign illustrates that menstruation remains controversial in advertising and public discourse.

There is a notable lack of research into how menstruation has been covered in Dutch newspapers, in particular, how it was represented historically. This gap in the literature highlights the importance and relevance of this thesis, which seeks to investigate historical newspaper articles to better understand how menstruation was discussed, or omitted, over time.

3 Related Works

This section reviews the technical challenges and approaches relevant to this thesis. The first two sub-research questions address technical challenges encountered during data acquisition and preparation. These include the quality of the Optical Character Recognition (OCR) in historical newspaper archives and the ambiguity of language that requires Word Sense Disambiguation (WSD). The latter subsections provide an overview of the Text Mining methods, specifically Topic Modeling and Sentiment Analysis, which are applied to answer the analytical research questions.

3.1 Optical Character Recognition

To digitally archive newspapers, Delpher uses Optical Character Recognition (OCR), a process by which printed or handwritten text is converted into machine-readable text [27]. The quality of OCR is influenced by several factors: the condition of the original material, the quality of the scan, and historical variations in Dutch spelling and typography. Due to wear and tear, suboptimal printing methods, and evolving spelling and grammar over the twentieth century, OCR output often contains errors.

Delpher is continuously working to improve its OCR technology and to correct historical inaccuracies. However, OCR errors can significantly impact text analysis, especially for unsupervised methods like Topic Modeling and Sentiment Analysis.

This study applies unsupervised learning methods, such as Topic modeling, which are sensitive to OCR errors. Studies by Walker et al. (2010, 2013) [40, 41] show that noise introduced by OCR errors has a more pronounced effect on unsupervised learning models than on supervised tasks, leading to challenges in identifying coherent topics and sentiments.

3.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is the task of determining the intended meaning of a word based on its context. This is particularly important when analyzing historical texts, where words may have different meanings than today or multiple senses within the same time period.

WSD approaches can be broadly categorized into two types [14], Machine Learning-based methods, which rely on labeled training data to learn patterns that distinguish between senses, and Dictionary-based methods, which identify word senses using predefined lexical resources. In this thesis, a machine-learning-based WSD model is applied to distinguish between relevant and irrelevant articles that include the use of ambiguous Dutch terms related to menstruation, such as *ongesteld*, which historically meant ‘unwell’ but now mainly refers to menstruation.

3.3 Text Mining

Text Mining allows large-scale analysis of textual data to identify patterns, themes, and sentiments. In this study, Artificial Intelligence methods are used to analyze Dutch newspapers at scale, as manual reading would be impractical.

To answer sub-question 3, *What topics related to the menstrual cycle were discussed in Dutch newspapers during the twentieth century, and how have these topics changed over the years?*, Topic Modeling is applied. To answer sub-question 4, *What sentiments did Dutch newspapers express on each topic, and how did these sentiments change over time?*, Sentiment Analysis is used.

3.3.1 Topic Modeling

Topic Modeling is an unsupervised machine learning technique that identifies recurring topics across a large collection of documents. Each document is represented as a distribution over topics, and each

topic is a distribution over words. One of the widely used Topic Modeling Methods is Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2001) [2].

Several studies have used Topic Modeling to analyze media and newspapers. Cho et al. (2019) [6] applied LDA to investigate trends in online newspaper articles about women’s health between 1993 and 2015. They segmented the dataset into four time periods and ran LDA separately on each subset, allowing them to analyze how topics shifted over time.

Dooremalen et al. (2021) [36] analyzed the coverage of 9/11 in newspapers from the United States, France, and the Netherlands between 2001 and 2015. They used LDA on sub-corpora by country and period, and validated topic intercoder agreement. When topics were unclear, they read the most representative articles to assign appropriate labels. This approach underscores the importance of human interpretation to validate results.

While many applications of Topic Modeling use data from short time spans, Hall et al. (2008) [17] applied LDA to decades of journal articles, identifying long-term thematic shifts. They employed post hoc analyses to track topic evolution.

Marshall (2013) [24] applied a more advanced variant, Correlated Topic Modeling (CTM), to British and French academic texts and British newspapers related to fertility. Unlike LDA, which assumes that topics are independent, CTM models correlations between topics. This is particularly useful when analyzing academic or news texts where co-occurring themes are common (e.g., health and politics). Marshall used CTM to track how topic interrelationships evolved over time, highlighting the flexibility of topic modeling for capturing both thematic content and structural shifts in discourse.

3.3.2 Sentiment Analysis

Sentiment Analysis (SA) is where people, opinions, assessments, attitudes, and emotions toward an individual, event, or topic are analyzed. Sentiment Analysis can be applied on three levels: feature, document, or sentence level. The goal of SA is to categorize the opinions into positive, neutral, or negative sentiments. The research on SA has expanded over the years, where research uses either Lexicon-based methods, Machine Learning-based, or a combination of the two [15].

The Lexicon-based methods are unsupervised, relying on predefined sentiment dictionaries, such as SentiWordNet or TF-IDF-weighted sentiment lexicons. These methods are relatively lightweight and do not require labeled training data. For example, Hossain et al (2021) [18] applied three lexicons (NRC, BING, and AFINN) to analyze the sentiment of newspaper headlines. In combination with word clustering, they determined the most frequent positive and negative words in the newspaper headlines.

In contrast, machine learning based methods use supervised algorithms, such as Support Vector Machines or Naive Bayes, trained on annotated datasets. In recent times, the emergence of transformer-based models, like BERT (Bidirectional Encoder Representations from Transformers) [13], has significantly advanced sentiment classification capabilities. BERT’s architecture has inspired various derivatives such as RoBERTa [23], DistilBERT [31], and ALBERT [22].

For the Dutch language, transformer models such as BERTje [13] and RobBERT [12] have been developed. RobBERT, created by Delobelle et al. (2020) [12], is specifically fine-tuned for sentiment

analysis on Dutch text and has shown strong performance on several NLP tasks.

De Bruyne et al. (2021) [11] investigated how well Dutch transformer models detect emotion in tweets and TV captions. They compared RobBERT and BERTje, and also explored hybrid methods that integrated lexicon-based features into the models. Their results showed that RobBERT consistently outperformed BERTje.

While transformer-based models show strong performance, Boukes et al. (2019) [3] cautions against the uncritical use of off-the-shelf sentiment tools. In a study of Dutch economic news, Boukes et al. found that generic sentiment analyzers were often unreliable and unsuitable for domain-specific tasks. The study highlights the importance of manual validation and the need for training sentiment tools for the specific language, genre, and context of the research.

4 Data

This section outlines the methods and results related to data selection, acquisition, preprocessing, and the solutions to the first two sub-research questions: (1) *To what extent does OCR quality affect the accurate retrieval of menstruation-related newspaper articles from the Delpher archive?* and (2) *How can semantic ambiguity in menstruation-related keywords (e.g., ongesteld, maandstonde, maandverband, tampon) be addressed in historical newspaper analysis?* It begins with the design decisions surrounding the dataset, discusses keyword selection and ambiguity, details metadata gathering and preprocessing, and concludes with a classification-based approach to disambiguate ambiguous terms.

4.1 Data Design

The dataset consists of Dutch newspaper articles published between 1900 and 1999, retrieved from the Delpher digital archive. Only documents categorized as *articles* were included, excluding advertisements, family notices, and image captions. Articles unrelated to journalistic reporting, such as trade records or shipping schedules, were removed.

In terms of geographical scope, Delpher distinguishes between newspapers published in different types of regions. These include National, Dutch East Indies/Indonesia, Netherlands Antilles, Regional/Local, Suriname, and the United States. This research only includes newspapers categorized as *National* or *Regional/Local* to ensure focus on publications from the mainland Netherlands.

While the dataset spans the entire twentieth century, Delpher’s coverage sharply declines after 1995 (Figure 1). This presents a limitation for analyzing discourse in the final years of the century.

As discussed in Section 2.3.3, World War II significantly disrupted newspaper publication due to censorship and propaganda. Delpher places a particular emphasis on archiving newspapers from this era, as can be seen in Figure 1.

To answer the last research question, the pillar affiliation of newspapers is needed. Since Delpher does not provide metadata on pillar affiliation, this information was collected manually through official newspaper websites, Wikipedia, or additional online sources. In some cases, the pillar could not be determined and was labeled ‘unknown’. These papers were often from small villages or towns. Some

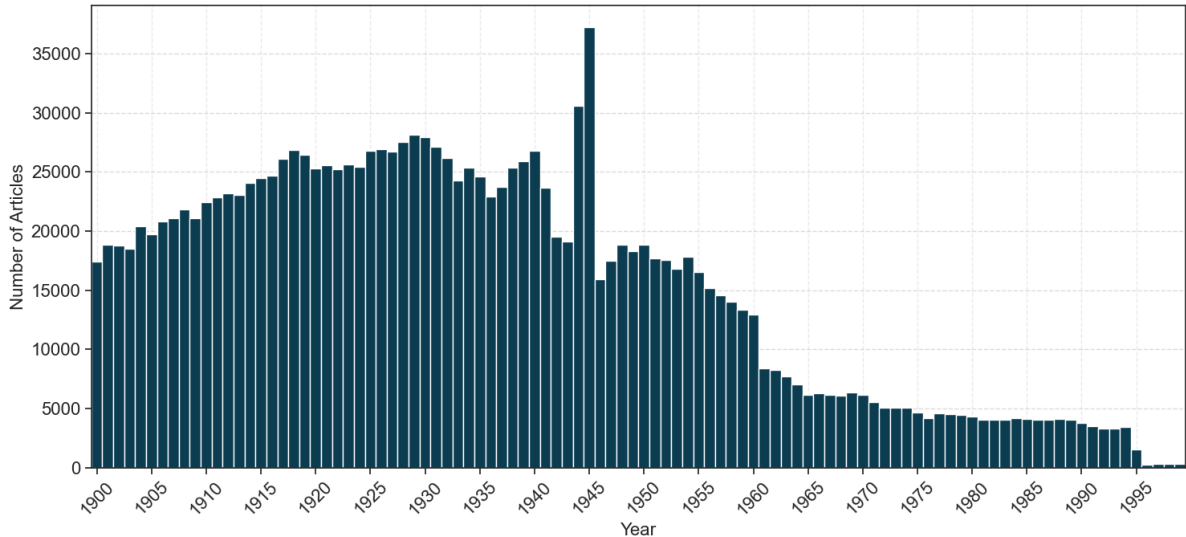


Figure 1: *The number of newspaper articles published between 1900 and 1999 archived on Delpher, the digital archive of the Dutch National Library.*

newspapers reflected ideologies beyond the four traditional pillars (e.g., Communist, Jewish, or anti-Semitic), and some were hybrid (e.g., Protestant/Liberal). For consistency, each newspaper is assigned its original pillar affiliation, acknowledging that ideological views may have shifted over time.

4.2 Data Acquisition

Articles were gathered using the Delpher API with guidance and tools provided by Delpher, including Jupyter Notebooks and Python documentation. The code is available through GitHub¹.

4.2.1 Keyword Selection

To gather the relevant newspaper articles from Delpher, keywords related to the menstrual cycle were used. The keywords were developed iteratively through domain knowledge and literature review (Table 1). The keywords used for the initial search were decided through personal knowledge of the menstrual cycle. By reading a random assortment of newspapers and through literary research, more keywords related to the menstrual cycle were discovered. The iterative process improved coverage and relevance.

4.2.2 Ambiguous Keywords and Contextual Challenges

This iterative process also revealed several challenges related to keyword ambiguity. Some keywords associated with the menstrual cycle also have alternative meanings, leading to large numbers of irrelevant search results.

The most problematic term was *ongesteld*. Currently, it is an euphemism for menstruation in Dutch (“*on her period*”), however, historically referred to general illness or feeling ‘unwell’, especially before the Second World War and continuing into the 1960s and 1970s. It was commonly used to describe anyone, regardless of gender, who felt unwell. For example, a newspaper article from January 2, 1900 (Figure 2)

¹<https://github.com/AHLebbink/AIThesisAHLebbink.git>

Dutch	English
menstruatie	menstruation
menstruatiecyclus	menstruation cyclus
menstruatiebloed	menstruationblood
menstruatiepijn	menstruation pain
menstruele	menstrual
menstrualecyclus	menstrual cycle
menopauze	menopause
menopauze	menopause
oestrogeen	estrogen
progesteron	progesterone
ovulatie	ovulation
menarche	menarche
dysmenorroe	dysmenorrhea
dysmenorrhea	dysmenorrhea
amenorroe	amenorrhea
amenorrhea	amenorrhea
amenorrhoea	amenorrhea
damesverband	feminine pads
wegwerpverband	disposable pads
inlegkruisje	panty liner
vrouwenkwaaltjes	female ailments
anticonceptiepill	birth control pill
Ambiguous Keywords	
ongesteld	menstruating
maandstonde	menstruation
maandverband	sanitary pads
tampon	tampon

Table 1: *Keywords used to find menstrual-related news articles.*

mentions the King of Belgium being *ongesteld* due to a foot injury from the previous winter. In rare contexts, *ongesteld* may also mean “not stated” or “unstated,” derived from the root *gesteld* (stated) with the negating prefix *on-*.

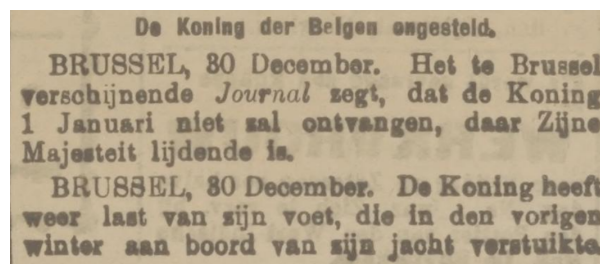


Figure 2: *Excerpt from a 1900 article describing the Belgian king as ongesteld due to a sprained foot.*

Figure 3 shows the frequency of the term *ongesteld* in Delpher articles over the century. Manual inspection of articles from the early and mid-20th century shows frequent reports of public figures (e.g., the pope, mayors, or royalty) being described as *ongesteld*. This was common at a time when newspapers were the primary source of updates about public figures, as radio and television were not yet widespread.

Additionally, OCR errors further complicated retrieval. For instance, *opgesteld* (“prepared” or “drawn up”) was frequently misread as *ongesteld*, leading to the inclusion of articles entirely unrelated to menstruation.

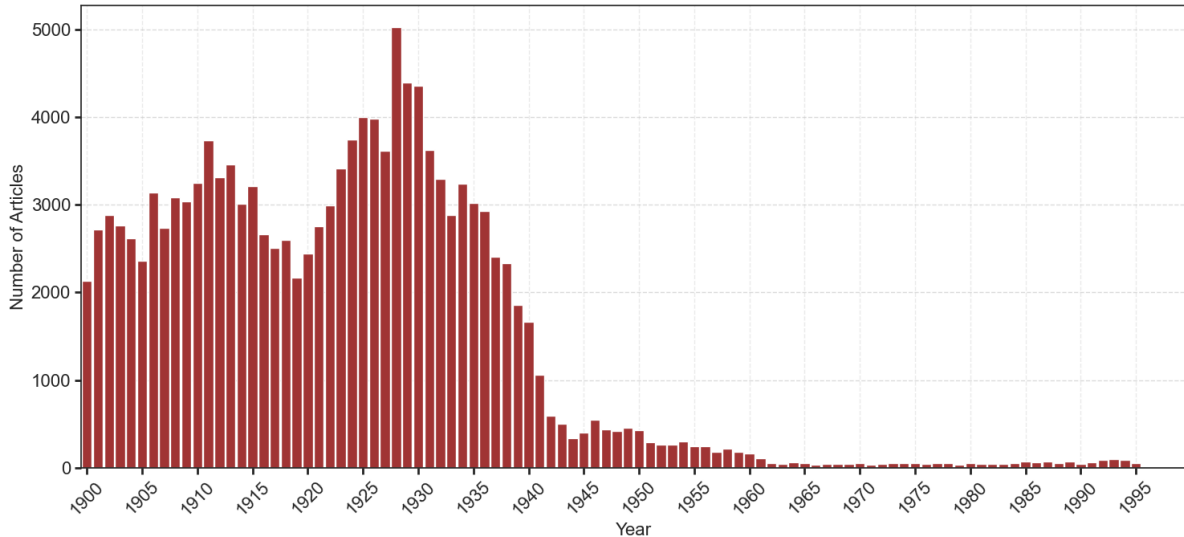


Figure 3: *Number of Delpher articles (1900–1999) containing the keyword ongesteld.*

Three other keywords also displayed ambiguity:

- ***Maandstonde***: While it refers to menstruation in old Dutch, it also means a memorial church service held one month after someone’s passing. These events were often printed in newspaper schedules, inflating irrelevant results.
- ***Maandverband***: In the context of menstruation, it means a sanitary napkin. However, it can also be interpreted as “monthly relation” or “monthly connection”.
- ***Tampon***: In Dutch, this word refers to the same menstrual product as in English. However, in French, it means “plug” or “buffer.” Because some Dutch newspapers occasionally included French-language content or reprinted international news, articles containing the word *tampon* were sometimes irrelevant. In some cases, French-language newspapers could be filtered out by title, but others required classification based on content.

A detailed explanation of how ambiguity was resolved is provided in Section 4.3 (Word Sense Disambiguation).

4.2.3 Article Collection and Filtering

Data was gathered in three stages: first, using non-ambiguous keywords (10094 newspaper articles), then with ambiguous terms (*maanstonde*, *maandverband*, *tampon*) (3148 articles), and finally, *ongesteld* data was gathered separately due to its high volume (134968 articles).

The OCR quality was an important aspect of this thesis. Delpher uses automated OCR techniques to digitize historical newspapers. However, the OCR accuracy is variable, especially for older materials with degraded typefaces, documents with visual noise, and irregular formatting.

To evaluate OCR reliability, a manual validation was conducted. A random sample of 76 articles was read in full to determine whether they were correctly retrieved and whether the OCR accurately found

the keyword(s). In 91% of the cases, the keyword appeared as expected, and the article was relevant. However, 4 out of 7 *ongesteld* cases involved OCR confusion between *ongesteld* (menstruating/unwell) and *opgesteld* (unstated), which altered the meaning entirely. This error arises because OCR struggles with character legibility, especially where font weight or smudging is present.

These findings show that while Delpher’s OCR is generally sufficient for the majority of the dataset, it is not fully accurate and it does introduce false positives and false negatives. False positives are when OCR mistakes produce incorrect word matches, while false negatives happen when true matches are missed entirely due to unrecognized or misread characters. While this study attempted to quantify and rectify false positives, the number of false negatives remains unknown due to time and hardware constraints.

To answer sub-question (5) *How do topics and sentiments on the menstrual cycle vary across newspapers with different ideological backgrounds (Protestant, Catholic, Socialist, Liberal) and geographical scopes (local/national)?* The affiliated pillar and geographical scope for each newspaper were needed. The pillars were acquired through manual research. The title of the newspaper that published an article and its geographical scope were gathered through a separate API request from the one that gathered the content of the newspaper. The result of each relevant newspaper with its affiliated pillar and publishing scope can be found in Appendix B.2.

4.3 Data Preprocessing

Before applying any analysis, the data went through many steps of selection, cleaning, and pre-processing. First, articles that were very short, under 200 characters (approximately 40 words), were removed, as they often lacked meaningful context or analytical information. Next, non-journalistic content such as trade schedules, Sinterklaas rhyming articles, and TV programme schedules were excluded. These were either done by the paper title, article title, content patterns, or metadata.

The remaining articles were then cleaned to remove the OCR errors. All the steps were done in Python. The data first went through a Dutch spellchecker using the Python package `Spellchecker`, followed by removing OCR artifacts, standardizing abbreviations, and stripping punctuation, symbols, digits, and standalone letters. Common city and country names were removed due to over-representation in the results of the topic model. For the same reason, honorifics (e.g., “Dr.”, “Mevr.”) were also removed. Furthermore, temporal expressions such as month and weekday names, including their abbreviations, were removed. At the end, the content went through the Dutch spellchecker again.

To handle the variation in article length, a segmentation approach was applied. Articles longer than 5000 characters (approximately 400 words) were truncated by selecting a window of 2500 characters before and after each menstrual-related keyword. If multiple keywords appeared close together, overlapping windows were merged to avoid duplication. Articles shorter than 500 characters remained unchanged.

Tokenization and lemmatization were performed using SpaCy’s Dutch-language model. The resulting data was lemmatized for Word Sense Disambiguation (WSD) and Topic Modeling tasks, while the original, non-lemmatized content with minimal stopword removal was used by Sentiment Analysis to preserve words with sentimental meanings.

After completing all preprocessing steps, the dataset included 108,876 articles containing the term *ongesteld*, 2624 articles with ambiguous keywords, and 9288 articles with other menstruation-related terms. After the preprocessing, only 9712 articles were identified as relevant. This filtering process is explained in more detail in the next section.

4.3.1 Word Sense Disambiguation

To determine whether articles found using ambiguous keywords truly relate to menstruation, a Word Sense Disambiguation (WSD) approach was used. A Logistic Regression model was chosen. The model was trained using labeled data, where each document (newspaper article) was annotated as ‘True’ based on the presence of non-ambiguous keywords. To reduce the class imbalance, non-ambiguous data was added to the ‘True’ class.

It is important to note that the labeled dataset used for training the classifier is not perfectly accurate. The purpose of the text classification step was to identify articles gathered by ambiguous keywords that do not contain other, more specific menstruation-related terms, but are still relevant to the topic of menstruation. Classification was used to separate truly relevant articles from unrelated ones that happened to include ambiguous keywords. This means a certain degree of noise was expected in the labels, but the model still helped refine the dataset to a much more precise subset for further analysis.

Through an exploratory and iterative process, initial results showed that *ongesteld* was used to refer to menstruation before the 1950’s, it only referred to being ‘unwell’ in the general sense. Therefore, only *ongesteld* articles published from 1950 onwards were included in the classification process. This historical context helped reduce the dataset (from 120788) to a more manageable size (to 16166) and improved the likelihood that ambiguous terms would occur in menstrual contexts. The majority of the *ongesteld* data was present in the earlier part of the 1950s, which is visualized in Figure 3.

The text classification pipeline used a TF-IDF vectorizer (n-gram range: 1-2, max features: 5000) followed by Logistic Regression. The model was trained using 80% of the dataset (n=12932) and tested on the remaining 20% (n=3234). The classifier achieved an F1-score of 0.94, with precisions of 0.92 for the negative class and 0.95 for the positive class.

To determine the correct threshold for classifying articles, a random sample of 30 articles from both datasets was evaluated by hand. Based on this evaluation, a Receiver Operating Characteristic (ROC) curve was generated Youden’s J statistic was used to identify the optimal threshold. For the *ongesteld* dataset, a threshold of 0.5 was chosen, which gave an F1-score of 0.774. Similarly, for the ambiguous dataset, a threshold of 0.5 gave the best performance, with an F1-score of 0.949.

From the ambiguous keyword dataset (excluding *ongesteld*), 349 of 2624 articles were classified as menstruation-related, of which 297 were labeled as false positives, meaning they did not contain non-ambiguous keywords but were classified as relevant by the model. For the *ongesteld* dataset (post-1950), 410 of 4254 articles were classified as relevant, 183 of which were also false positives. These false positives are important: they represent articles that would have been missed by keyword matching alone but were successfully retrieved through WSD.

When combined with the non-ambiguous keyword dataset and removing duplicates, this resulted in a

final set of 9712 relevant newspaper articles. The frequency of articles per year is visualized in Figure 4.

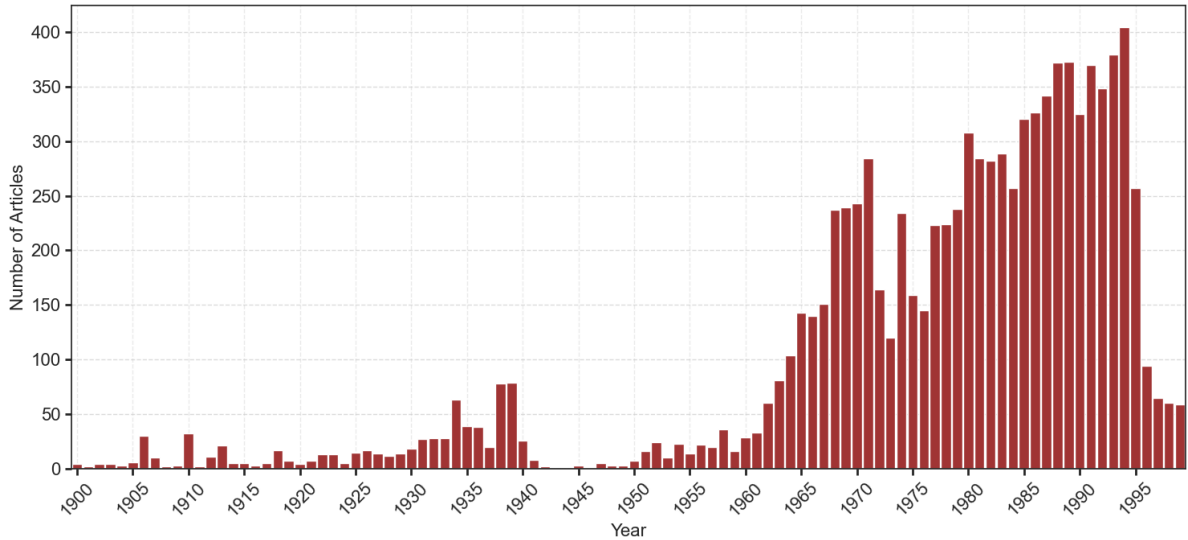


Figure 4: *The total number of relevant newspaper articles per year.*

5 Topic Modeling

The menstrual cycle is associated with a wide range of societal, biological, and medical discourses. To determine which topics are discussed in Dutch newspaper articles, topic modeling was applied to the corpus. The following section addresses the research question (3) *What topics related to the menstrual cycle were discussed in Dutch newspapers during the twentieth century, and how have these topics changed over the years?*

5.1 Method

Latent Dirichlet Allocation (LDA) [2] was chosen as the topic modeling algorithm. The implementation used was `LdaModel` from the Python library `Gensim` [30]. Each newspaper article was treated as an individual document.

Before applying LDA, the data was tokenized, lemmatized, truncated around the keywords, and stop words and function words were removed. Words that appeared in fewer than 5 documents or in more than 10% of the corpus were removed to reduce noise and over-represented terms. These parameters were refined by looking at the repeats in the words of the topic, for example, when no words or noise were removed, words like ‘menstruation’ and ‘woman’ were in almost every topic, due to the over-representation of these words in the corpus. By removing these, more refined topics could be found in the Dutch newspapers.

The `LdaModel` was trained with 50 passes. Both the alpha and eta parameters were set to ‘auto’, enabling the model to learn asymmetric priors for document-topic and topic-word distributions, respectively. All other parameters remained at their default settings. A `random_state` of 42 was used to ensure reproducibility.

The number of topics was determined manually through iterative testing, varying the topic count between 5 and 25. By connecting the resulting topic words to the newspapers assigned to the topic, at each variant of the results, the different types of topics were interpreted. Below 15 topics, meaningful nuances were lost, too many topics were combined, and above 20 topics, there were too many topics with similar meanings, and topics that were assigned only a handful of articles. This finally resulted in 17 topics.

To assign a dominant topic to each article, the topic with the highest probability was selected. At first, to interpret and label each topic, the top 10 keywords per topic were used to assign an initial label. Then these names were refined by reading articles assigned to a topic. The articles with the highest assigned probability for that topic, and a set of 5 random articles. Through this, the name and meaning of the topic were refined.

This combination of statistical modeling and qualitative review ensured that topics were not only computationally distinct but also semantically meaningful within the historical and cultural context.

To visualize the temporal trend of the topics, the Python packages `matplotlib` and `seaborn` were used. A graph was generated that depicts the number of articles per topic every five years. To illustrate the total number of articles for each five-year period, a histogram is plotted behind it with its own separate y-axis.

5.2 Results and Discussion

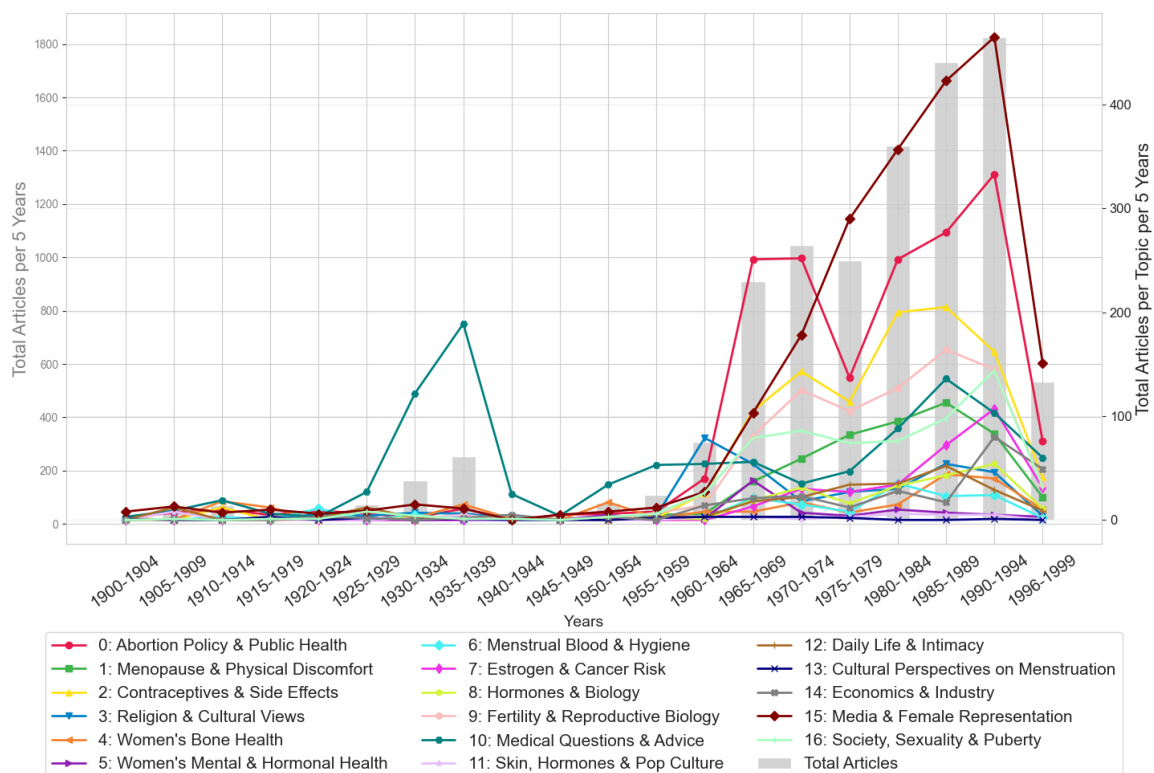


Figure 5: The frequency of newspaper articles per topic every five years. The bar graph depicts the total number of articles published in each of the 5 years.

Topic	Keywords	Topic Name	#
15	story, movie, thing, all, piece, look, photo, beautiful, full, lady	Media & Female Representation	2098
0	abortion, minister, public health, committee, advice, yesterday, law, ban, pay, hospital	Abortion Policy & Public Health	1688
10	blood, skin, surgery, eating, pain, answer, condition, breast, symptom, advice	Medical Questions & Advice	1073
2	side effect, contraception, Haspels ² , market, function, IUD, condom, take, pharmaceutical, method	Contraceptives & Side Effects	1011
9	egg cell, uterus, ovulation, method, fertilization, cycle, ovulation, sperm, fertilized, gynecologist	Fertility & Reproductive Biology	809
16	sexual, parent, boy, school, father, family, marriage, relationship, daughter, sexuality	Society, Sexuality & Puberty	641
1	menopause, physical, pain, psychological, phenomenon, feeling, change, headache, all kinds, migraine	Menopause & Physical Discomfort	508
3	church, catholic, century, god, appear, pope, culture, consider, writer, subject	Religion & Cultural Views	344
7	estrogens, breast cancer, estrogen, menopause, risk, heart, cancer, effect, estrogen, vascular disease	Estrogen & Cancer Risk	315
14	million, guilders, company, market, product, rise, price, average, US ³ , billion	Economics & Industry	260
12	eat, bed, little, house, half, stop, talk, night, home, have	Daily Life & Intimacy	236
4	osteoporosis, bone, vitamin, dissertation, calcium, anorexia, band-aid, fat, hip, Enk ⁴	Women's Bone Health	218
8	ovary, estrogen, male, produce, cell, progesterone, animal, brain, pituitary, blood	Hormones & Biology	218
6	blood, tampon, animal, meat, plant, iron, cross, bacteria, slip, discharge	Menstrual Blood & Hygiene	153
5	syndrome, depression, Defares ⁵ , scientific, article, childbirth, vascular disease, study, heart attack, hormonal	Women's Mental & Hormonal Health	80
11	acne, male, dog, cat, effect, Groeneveld ⁶ , feature film, Alberda ⁷ , anabolic, skin	Skin, Hormones & Pop Culture	39
13	modern, traditional, population, menopause, experience, Israel, city, China, Chinese, positive	Cultural Perspectives on Menstruation	21

Table 2: *Menstruation-related topics in Dutch newspaper articles from 1900 to 1999.*

²Prof. dr. Arie Haspels, inventor of the morning-after pill and co-developer of the abortion pill.

³US: United States (of America).

⁴Dr. Adam van Enk, a gynecologist from Amsterdam.

⁵James George Defares, in 1967, wrote about the influence of the contraceptive pill on aging.

⁶General Practitioner Dr. Frans Groeneveld.

⁷Dr. Albertus Alberda, gynecologist Dijkzigt Hospital Rotterdam.

The LDA model produces 17 distinct topics. Table 2 shows each topic's ID, top ten keywords, assigned label, and total number of articles. The original Dutch topic words are included in Appendix A. The distribution of topics over time is visualized in Figure 5.

Overall, topic frequencies roughly correlate with the overall number of articles published over time that relate to menstruation, but there are several trends and interesting patterns that emerge.

Topic 15 (Media & Female Representation) was the most frequently occurring topic, appearing in 2098 articles. The top ten words associated with this topic suggest a theme of media, femininity, and the objectification of women in pop culture and media. Reading a random sample of articles confirmed this impression. One article from September 1990, for instance, was about Dutch women in Hollywood, reflecting on how female identity and appearance were portrayed in the media. Another article, a book review, focuses on a girl growing up in 1861. She is held back by traditions, her path is dictated by her menstrual cycle, and the societal expectations to marry and have children, which prevent her from having an education or a career. The review highlights the pressure and expectations women face in their lives. Interestingly, this topic also includes a recurring genre of article from the 1950s: "lost and found" notices. These often list women's purses containing items such as 'damesverband' (sanitary pads), and although seeming different from the topic, it was assigned due to shared vocabulary.

Figure 5 shows the number of articles per topic every five years, next to the total of articles per five years. As expected, the number of topic-assigned articles largely follows the general increase in published articles across the century. However, some topics rise or decline independently of the overall pattern. The steep drop in the final five years is due to the limited number of articles archived by Delpher post-1995, possibly a result of the digital transition to online media.

The frequency of Topic 15 follows the general increase in articles about the menstrual cycle. Given the variety of themes it encompasses, femininity in media and lost and found listings, it is unsurprising that this topic grew alongside the overall corpus.

Topic 0 (Abortion Policy & Public Health) is the second most occurring topic, assigned to 1688 articles. Its occurrence over the century aligns closely with national debates on reproductive rights. Abortion was legalized in the Netherlands in 1984, but this newspaper coverage can be traced back to the 1960s. The coverage during the sixties and seventies can be connected to the second-wave feminism. The drop in the late seventies can be attributed to the overall drop in newspaper publications about the menstrual cycle during that period.

Topic 10 (Medical Conditions and Surgery), with 1073 articles, is the third most frequent topic. The topic peaks in the 1930s, tapers off mid-century, and then resurges in the 1980s. Articles typically contained responses to health-related questions submitted by readers. Examples from the therapies include discussion on childbirth dates and epilepsy. Menstruation was either mentioned directly or as a hormonal factor contributing to other medical ailments. These columns changed in title from "*Medical Questions*" in the 1930s to "*Just Ask*" in the 1950s, and later "*Doctor, What Do You Think?*" in the 1960s. Even in the 80s, similar content was still published in newspapers, answering questions about the menstrual cycle. More in this topic is discussed in Chapter 7.2.2.

Topic 2 (Contraceptives & Side Effects) became prominent after the 1960s. This topic covers the contraceptive pill and the discussion surrounding it, reflecting the societal and medical impact when it was introduced in the Netherlands in 1962. Notably, the term “Haspels” refers to Prof. Dr. Arie Andries Haspels, a Dutch gynecologist and professor in obstetrics. Together with his team, he developed the morning-after pill, and he was a major advocate of contraceptives. His presence reflects how newspapers shaped discourse around reproductive health.

Topic 9 (Fertility & Reproductive Biology) is represented by 809 articles, which focus on the news on fertilization technology, such as IVF. The topic’s rise in the 1970s aligns with the major developments in reproductive medicine; most interesting is the birth of the first IVF baby in 1978.

Topic 16 (Society, Sexuality & Puberty) includes 641 articles and is about how menstruation intersects with broader everyday themes. The articles often contain news on the changes in average height and weight of society, the onset of puberty, and family dynamics. The topic rose in the late sixties and onward, overlapping with the sexual revolution and changing norms surrounding adolescents’ health and sexuality.

Topic 1 (Menopause & Physical Discomfort) covers the physiological and psychological effects of menopause, including pain, migraines, and mood changes. With 508 articles, and a rise in frequency mirroring the overall articles. The occurrence suggests that menopause and its symptoms were recognized by society to a certain degree, and it was a recurring theme in the latter half of the twentieth century.

Topic 3 (Religion & Cultural Views) is about how religions and the Catholic Church view topics surrounding the menstrual cycle. For example, the view of the Pope towards the contraceptive pill. This topic is the most frequent one in the early 1960s, when the contraceptive pill came over from the USA in 1963. One article published church news in 1963 on how the bishops of the Roman Catholic church discussed marital problems and Roman Catholic morals, where they were concerned about birth control, like the contraceptive pill. Another column published in *De Volkskrant* in 1986 reports on how a cardinal of a Catholic church has said that women do not belong in the church.

Topic 7 (Estrogen & Cancer Risk) is about articles that discuss the effect of estrogen in the female body, from where it makes a woman from a girl, how the deterioration of the hormone occurs after menopause, and how research into how this hormone affects cancer. over the years, this topic has had a small presence, with a rising number of articles starting in the seventies, and peaking in the early nineties, following the rise of scientific progress.

Topic 14 (Economy & Industry), though more distantly related to menstruation, touches on the commercial and financial aspects of menstrual health. Articles discuss prices of sanitary and medical products, including contraceptives. Companies like AkzoNobel, involved in contraceptive production, were featured in stock exchange reporting during the nineties.

Topic 12 (Daily Life & Intimacy) does not have distinct peaks during the twentieth century; the number of articles per year rises with the total articles found per year. The topic words show a less clear topic related to menstruation, analyzing its assigned articles, discovers that it is about the menstrual cycle has effects on daily life, eating patterns, family, pregnancy, friendships, and relationships.

Topic 4 (Women’s Bone Health) has 218 articles assigned to it. The words of this topic also contain

a name, Dr. Adam van Enk, who is a gynecologist from Amsterdam. He is mentioned multiple times in the articles, due to his commenting on how women's bone health deteriorates over the years, and how there is not enough research on what the true reason is for more broken hips in elderly women.

Topic 8 (Hormones & Biology) is about human biology, the influence of hormones, and how they work differently in the male and female body. It contains the word animal, due to the research done on hormones on animals. The results of this research are reported on in newspapers.

Articles that are assigned topic 6 (Menstrual Blood & Hygiene) are often on the discussion surrounding sanitary products, bacteria, and public cleanliness. These are in combination with product commercialization and evolving health standards. This topic is not frequently represented and has only a small peak around the early eighties.

Topic 5 (Women's Mental & Hormonal Research) articles present the research done surrounding women's health, like depression, childbirth, vascular diseases, and heart attacks. One of the words in this topic is the name Defares, which is the last name of James George Defares, a Dutch physician, physiologist, and author. In 1967, he wrote to the editors of the Netherlands Journal of Medicine on the subject of 'the contraceptive pill and aging'. He wrote his idea about how the contraceptive pill could lead to accelerated biological aging and a heightened risk of vascular diseases. Defares and his opinions were in newspapers in 1967. However, the academic world did not agree, and quite quickly after this, he was forbidden to teach medical students, and in 2001, he was chosen as the sixth of the top 20 charlatans of the twentieth century by the Dutch *Association Against Quackery*. Topic 5 had a peak in the late sixties, due to 1967 being the publishing year of the book that Defares wrote.

Topic 11 (Skin, Hormones & Pop Culture) is about appearance, dermatology, and celebrity. Frans Groeneveld, a family doctor, comments in the news articles on his research on menopause and how it is seen as a negative thing. This topic has a small presence but is indicative of a broader cultural narrative linking hormones with beauty and aging.

Topic 13 (Cultural Perspectives on Menstruation), the least frequent topic, with only 21 articles, covers mostly the perspective of non-Western cultures, including China and Israel, on the menstrual cycle.

These seventeen topics represent the themes discussed in newspaper articles that cover menstruation-related themes.

5.3 Discussion

The use of topic modeling has shown the multiple ways menstruation was discussed in Dutch newspapers throughout the twentieth century. Menstruation itself was rarely discussed on its own in newspapers; it often intersected with a wide range of societal, political, medical, or cultural topics. The high frequency of Topic 15 (Media & Female Representation), which includes articles about women in film, beauty, and lost and found reports, is very interesting. It demonstrates how menstruation could be indirectly mentioned in connection with stereotypical representations of femininity and womanhood. Other prominent topics, such as *Abortion policy* and *Contraceptives & Side Effects*, reflect key social and political changes in society. Their temporal spikes correspond with significant historical events, such as the introduction of

the contraceptive pill and the legalization of abortion. These patterns show that menstruation was more visible in public discourse during moments of political tension or medical advancements, especially in debates around reproductive autonomy.

The appearance of the topics *Medical Questions & Advice* and *Estrogen & Cancer Risk* also highlights how newspapers functioned as a source of health education, both for general readers and for women trying to understand their bodies. These topics are present throughout the decades, indicating that menstruation-related issues were relevant to readers over time, even if they were not always framed as such.

Less frequent but still meaningful topics, like *Religion & Cultural Views*, *Menstrual Blood & Hygiene*, and *Cultural Perspectives on Menstruation*, show that cultural and religious views occasionally entered the public conversation. In these instances, menstruation was often tied to moral or ritual discourse.

Overall, the topics demonstrate that menstruation was not a taboo in the sense of being completely absent; it was often mentioned indirectly or in relation to more ‘acceptable’ topics like medicine or public policy. The fact that the personal or bodily topics (e.g., hygiene, blood, mental health) were the least frequent suggests that the more intimate parts of menstruation were less often addressed in public media, or that they were framed in a more medical or euphemistic language.

Topic modeling has shown that menstruation-related discourse in newspapers varied, shaped by medical, political, and cultural forces. The topic was not hidden, but also rarely the center of an article.

6 Sentiment Analysis

This section addresses the sentiment depicted in newspaper articles discussing the menstrual cycle. It answers research question (4) *What sentiments did Dutch newspapers express on each topic, and how did these sentiments change over time?*

6.1 Method

Sentiment analysis was applied to identify the emotional tone of newspaper articles related to menstruation. The goal was to determine whether the overall sentiment of an article was positive, neutral, or negative.

To accomplish this, the Dutch-language transformer model RobBERT [12] was used, which is a fine-tuned version of RoBERTa [23] specifically adapted for Dutch sentiment classification tasks. The model used was `DTAI-KULeuven/robbert-v2-dutch-sentiment`, implemented using HuggingFace’s `RobertaForSequenceClassification` and `RobertaTokenizer`.

However, this pre-trained model was originally trained on Dutch book reviews from the 21st century, which makes it less suited for analyzing twentieth-century newspaper articles, especially those that focus on a specific topic like menstruation. Therefore, fine-tuning the model on a more representative dataset was necessary.

To make such a dataset, 1000 articles were randomly selected from the complete corpus (`random_state = 42`) and were labeled using GPT-4o. Each article was categorized as expressing a *positive*, *neutral*, or

negative sentiment. This labeling process resulted in 556 neutral, 392 positive, and 52 negative articles.

This labeled subset was then used to fine-tune the RobBERT model, with a train-test split of 80/20%. During preprocessing, function words and stop words were not removed to preserve semantic meaning. The model was fine-tuned using a learning rate of $2e-5$, and a batch size of 8 for both training and evaluation. Training was performed for 15 epochs, with a weight decay rate of 0.01 to avoid overfitting.

The evaluation metrics of the fine-tuned model are reported in Table 3.

Class	Precision	Recall	F1-score	Support
Negative	0.00	0.00	0.00	9
Neutral	0.82	0.77	0.79	117
Positive	0.81	0.70	0.75	74
Accuracy			0.71	200
Macro Avg	0.54	0.49	0.52	200
Weighted Avg	0.78	0.71	0.74	200

Table 3: *Evaluation metrics of the fine-tuned RobBERT sentiment model.*

Due to the limited number of negative articles in the test set (only 9), and only 43 articles in the training set, the model was unable to correctly classify any of them, resulting in an F1-score of 0.00. However, performance was stronger for the neutral and positive classes, achieving F1-scores of 0.79 and 0.75, respectively. This low performance was due to both a low representation in negative cases and a time limit. Fine-tuning of the model took longer than expected, and further training was not a possibility.

6.2 Results & Discussion

The sentiment classification for the 9712 articles was as follows: 7419 (76%) neutral, 1939 (20%) positive, and 354 (4%) negative. This indicates that the majority of articles maintained a neutral tone when addressing menstruation, which is consistent with the formal and informative tone of journalistic writing.

The relatively low number of negative articles might seem surprising, especially considering the historical stigma and taboo associated with menstruation. One might expect a higher number of negative representations, whether in the form of discriminatory social rules, religious restrictions, or the medicalization of menstruation as an illness. Additionally, menstruation is often associated with discomfort, pain, and inconvenience. In modern times, when individuals speak about menstruation on social media, the tone is often overwhelmingly negative, often in relation to their own negative experiences [10, 35].

However, newspaper articles tend to present menstruation from a much more distant and impersonal perspective. By examining a random set of articles, it was discovered that descriptions are often more factual, and even statements with more negative emotions are rarely expressed in such a way. This likely explains the low proportion of articles classified as negative.

Figure 6 shows the sentiment distribution over time for every five years. Neutral sentiment remains dominant throughout the century, while positive sentiment is fairly consistent but less frequent. Negative sentiment only began to appear after the overall rise in article volume in the second half of the twentieth century.

To better understand the content and framing behind each sentiment label, the distribution of the sentiments across the 17 topics was analyzed. Table 4 shows how each topic breaks down in terms of

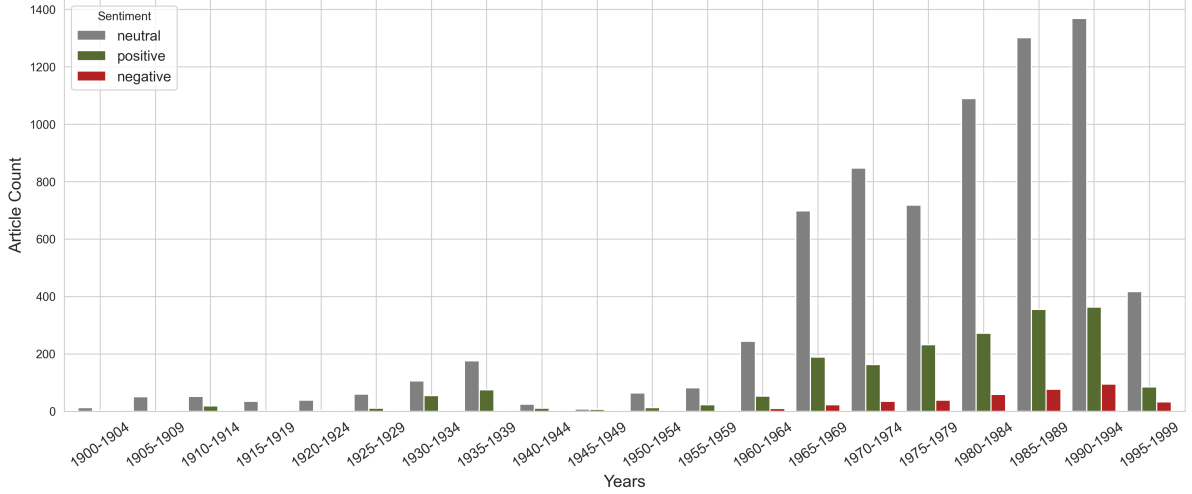


Figure 6: *The sentiment distribution over each 5-year period.*

positive, neutral, and negative sentiment. For example, Topic 15 (Media & Female Representation) includes the largest number of negative articles (133), possibly due to portrayals of gender stereotypes, objectification, or social pressure. Topic 0 (Abortion Policy & Public Health) also contains a notable share of negative sentiment, which aligns with the emotionally charged nature of abortion debates. Articles that are assigned Topic 7 are proportionally assigned more negative sentiments compared to topics with the same number. This makes sense due to the topic being about cancers, which is a negative topic in itself.

Sentiment		Absolute			Proportional		
Topic	Name	Neutral	Pos	Neg	Neutral	Pos	Neg
15	Media & Female Representation	1560	405	133	0.74	0.19	0.06
0	Abortion Policy & Public Health	1381	261	46	0.82	0.15	0.03
10	Medical Questions & Advice	787	255	31	0.73	0.24	0.03
2	Contraceptives & Side Effects	874	110	27	0.86	0.11	0.03
9	Fertility & Reproductive Biology	499	294	16	0.62	0.36	0.02
16	Society, Sexuality & Puberty	316	296	29	0.49	0.46	0.05
1	Menopause & Physical Discomfort	406	84	18	0.8	0.17	0.04
3	Religion & Cultural Views	286	47	11	0.83	0.14	0.03
7	Estrogen & Cancer Risk	263	29	23	0.83	0.09	0.07
14	Economics & Industry	245	5	10	0.94	0.02	0.04
12	Daily Life & Intimacy	152	67	17	0.64	0.28	0.07
4	Women's Bone Health	191	21	6	0.88	0.1	0.03
8	Hormones & Biology	185	27	6	0.85	0.12	0.03
6	Menstrual Blood & Hygiene	131	16	6	0.86	0.1	0.04
5	Women's Mental & Hormonal Health	65	12	3	0.81	0.15	0.04
11	Skin, Hormones & Pop Culture	36	3	0	0.92	0.08	0.0
13	Cultural Perspectives on Menstruation	16	3	2	0.76	0.14	0.1

Table 4: *Sentiment distribution across topics.*

To better interpret the relationship between sentiment and topics over time, the articles were filtered by sentiment category and plotted in three separate graphs, each showing the topic distribution across five-year intervals.

Figure 7 shows the number of positively classified articles assigned to each topic. The most frequent topics (Topic 15 (Media & Female Representation), Topic 0 (Abortion Policy & Public Health), and Topic 10 (Medical Questions & Advice)) remain prominent. However, two additional topics stand out more clearly in the positive sentiment category compared to the overall dataset. First, Topic 9 (Fertility & Reproductive Biology) shows a notable increase of positive articles in the seventies and eighties. This overlaps with significant advances in reproductive technology, such as the development and public discussion of IVF, including the first IVF baby in 1978. These breakthroughs were often reported in newspaper articles with an optimistic framing, which aligns with the positive sentiment label. Second, Topic 16 (Society, Sexuality & Puberty) also displays heightened variation in positive sentiment across time. This fluctuation may reflect the evolving opinion on sexuality and gender roles that started in the 1960s, with the start of second wave feminism in the Netherlands.

Figure 8 presents the topic distribution for articles labeled as neutral. Here, the distribution largely mirrors the overall topic frequency patterns observed in the complete dataset. This reinforces the earlier conclusion that most newspaper coverage of menstruation took on a factual, informative tone.

In contrast, Figure 9 conveys a much different distribution for negatively classified articles. The majority of the articles fall under Topic 15. This suggests that negative sentiment most frequently occurs in media portrayals of women’s discussions surrounding societal expectations of femininity.

6.2.1 Discussion

The sentiment analysis adds another layer to the understanding of how menstruation was represented in twentieth-century Dutch newspapers. The large number of neutral sentiments suggests that menstruation was often framed in a factual, depersonalized way. Aligning with the journalistic role newspapers play, informative rather than expressive. The more positive tones the topic *Fertility & Reproductive Biology* highlights moments where menstruation intersects with broader themes of medical innovation and social progress. In contrast, negative sentiments were rare but were concentrated around medical news, such as cancer and the portrayal of women in the media.

The lack of negative-labeled training samples resulted in a weakened model performance in detecting negative sentiments. Additionally, the sentiment classification mode, while state-of-the-art for Dutch, was trained and fine-tuned within time and computational constraints. Future improvements could include a larger and more balanced dataset. further fine-tuning with domain-specific data.

7 Pillarization & Geographical Scope

To understand how regional and ideological orientations shaped newspaper discussions on menstruation, this section explores how topics and sentiments varied across the pillars of Dutch society and between national and local newspapers by answering research question (5) *How do topics and sentiments on*

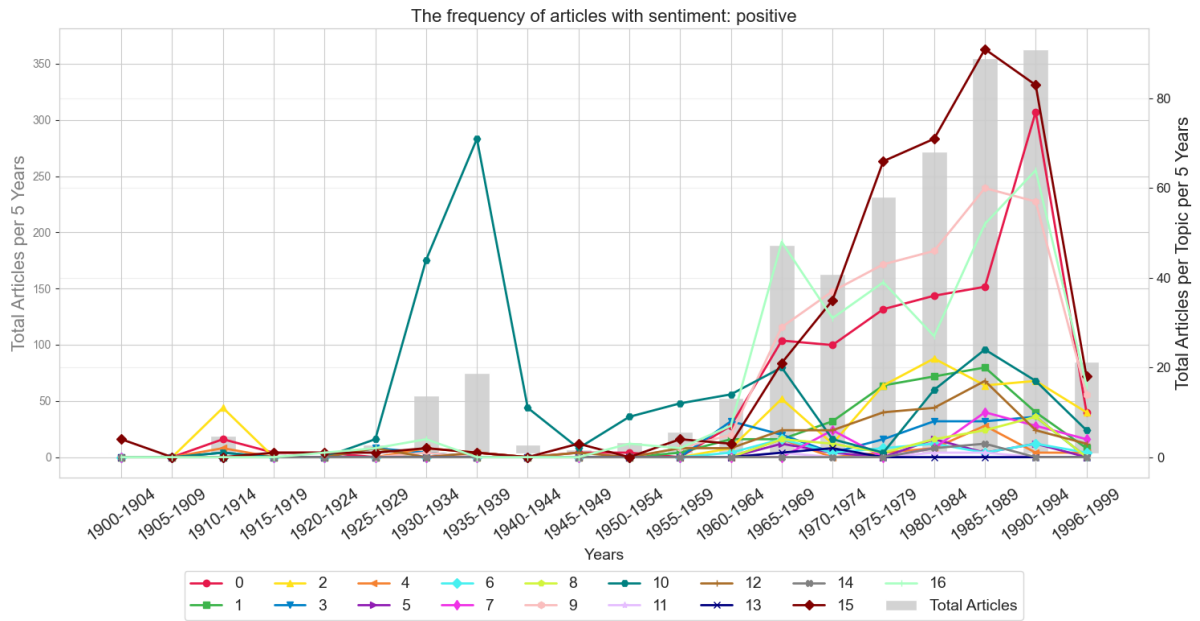


Figure 7: *The topic distribution of articles with a positive sentiment.*

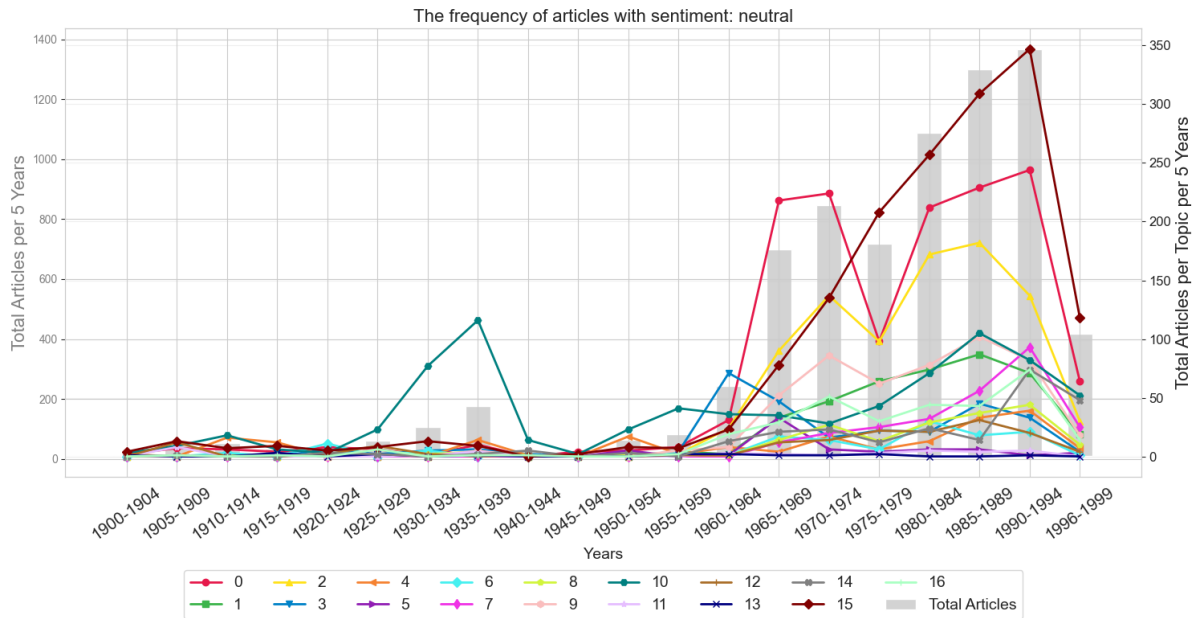


Figure 8: *The topic distribution of articles with a neutral sentiment.*

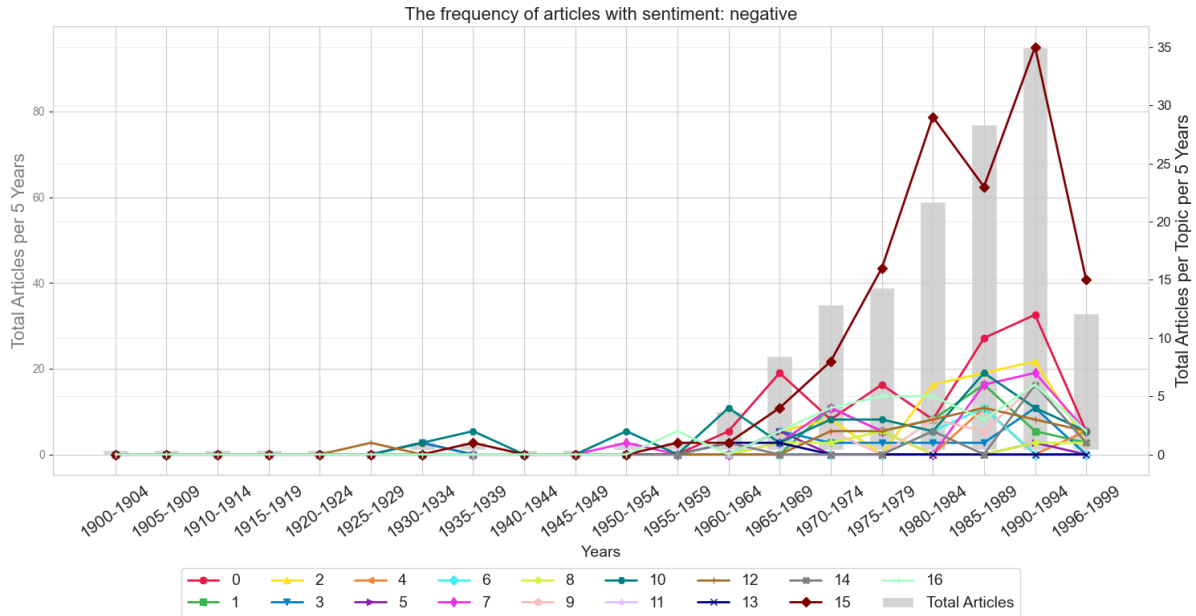


Figure 9: The topic distribution of articles with a negative sentiment.

the menstrual cycle vary across newspapers with different ideological backgrounds (Protestant, Roman Catholic, Socialist, Liberal) and geographical scopes (local/national)?

7.1 Method

Pillarization (*verzuiling*) refers to the segmentation of Dutch society into distinct ideological and religious pillars, most prominently Protestant, Roman Catholic, Social Democratic, and Liberal.

Using metadata from the Delpher archive, newspapers were classified by geographical scope (national vs local/regional), while pillar affiliations were gathered through historical sources. Each newspaper article in the dataset was assigned to a pillar and scope, allowing the results of the topic modeling and sentiment analysis to be further segmented and interpreted.

7.2 Results

The fifteen newspapers that most frequently published menstruation-related articles during the twentieth century are listed in Table 5. These represent a broad spectrum of Dutch society, including national daily newspapers like *Algemeen Dagblad*, *De Volkskrant*, and *Trouw*, and local newspapers like *Leeuwarder Courant* and *Limburgsch Dagblad*. Together, the 15 newspapers contribute to 90% of the total dataset, while the remaining 107 newspapers each contributed 1.0% or less. The complete results are represented in Appendix B.2. The geographical scope of newspapers is skewed toward national publications. Of the 9712 articles analyzed, 7375 (~76%) were published by national newspapers, while 2337 (~24%) came from regional or local outlets. This shows that menstruation and related health issues were more frequently discussed in national conversations, and not a topic often discussed by local newspapers.

The four most frequent pillars represented in the total dataset are Social Democratic (2800), Liberal (2773), Protestant Reformed (1793), and then Roman Catholic (1388). These four pillars account for

Newspaper	Frequency	Percentage	Geographical Scope	Pillar
Algemeen Dagblad	996	10.2%	National	Liberal
De Volkskrant	910	9.4%	National	Roman Catholic
Het Parool	890	9.2%	National	Social Democratic
Het Vrije Volk	818	8.4%	National	Social Democratic
Trouw	791	8.1%	National	Protestant Reformed
De Telegraaf	746	7.7%	National	Liberal
NRC Handelsblad	725	7.5%	National	Liberal
Nieuwsblad van het Noorden	667	6.9%	Local	Social Democratic
Leeuwarder Courant	626	6.4%	Local	Protestant Reformed
Limburgsch Dagblad	383	3.8%	Local	Roman Catholic
Nederlands Dagblad	353	3.6%	National	Protestant Reformed
De Waarheid	270	2.8%	Local	Communist
De Tijd	224	2.3%	National	Roman Catholic
Het Volk	218	2.2%	National	Social Democratic
Algemeen Handelsblad	144	1.5%	National	Liberal

Table 5: *The fifteen most frequent newspapers that published menstrual-related articles between 1900 and 1999.*

94% of all articles. The rest, including Communist, Neutral, and Jewish papers, are represented in Table 6.

7.2.1 Scope and Pillars

The distribution of pillar affiliations across the geographical scope shows interesting patterns, as presented in Table 7. All four major pillars (Social Democratic, Liberal, Protestant, and Roman Catholic) have a larger presence in national newspapers, but the degree varies greatly. The table also shows its proportional distribution.

Most interesting is the Liberal pillar, which is mostly represented by national media, with 2777 liberal articles; 98% were published by national newspapers. This likely connects to the large nationally distributed outlets such as *Algemeen Dagblad*, *De Telegraaf*, and *NRC Handelsblad*. This low frequency in local Liberal newspapers can be attributed to there being less of, being less archived on Delpher, or them printing less on the topic of menstruation, but this seems less likely when comparing it to the amount of other local articles in other pillars.

A similar pattern can be seen for the Communist newspapers, where almost all articles originated from national publications. This is mostly due to the national newspaper *De Waarheid* (translated: The Truth) that published 270 of the 298 articles.

The other three major pillars, Social Democratic, Protestant, and Roman Catholic, also show a national bias, but to a lesser degree. These pillars were more evenly distributed across national and local platforms.

Interestingly, nearly all articles published by newspapers with an unknown pillar affiliation came from local sources. This may be due to that many local newspapers, particularly smaller or short-lived ones,

Pillar	Frequency	Percentage
Social Democratic	2800	28.83%
Liberal	2777	28.59%
Protestant Reformed	1801	18.54%
Roman Catholic	1771	18.24%
Communist	305	3.14%
Unknown	88	0.91%
Neutral	66	0.68%
Jewish	39	0.40%
Christian	22	0.23%
Center-Right	16	0.16%
Anti-Revolutionaries	14	0.14%
Patriotic	5	0.05%
Independent	3	0.03%
Conservative	3	0.03%
National Socialist	1	0.01%
Anti-Semitic	1	0.01%

Table 6: *The frequency of articles distributed over the pillars.*

did not have a clear political or religious alignment, or have since lost documentation of this information. In these cases, classification is either ambiguous or not applicable, especially for newspapers that focused primarily on local events, family announcements, and community life rather than political discourse. In comparison to this, the neutral papers were mostly local to, but unlike unknown papers, it was clear that they did not fall under a specific pillar.

Finally, the fringe pillars (e.g., Jewish, Center-Right, Anti-Revolutionary, or National Socialist) were in such low numbers, both because the number of newspapers representing these pillars was low, and also shows that these newspapers did not place their focus on discussions surrounding menstruation. Overall, this shows that national newspapers dominated menstruation-related coverage across most ideological lines.

7.2.2 Topic Modeling

The distribution of the 4 major pillars and the topics is shown in Figure 10, and the full distribution can be found in Appendix B.1.

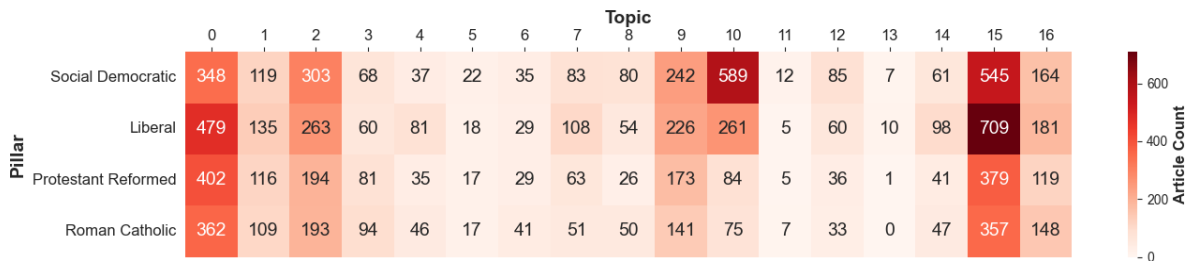


Figure 10: *Heatmap of the number of articles per pillar for every topic.*

Geographical Scope	Absolute		Proportional	
Pillar	National	Local	National	Local
Social Democratic	1946	854	0.7	0.3
Liberal	2710	67	0.98	0.02
Protestant Reformed	1154	647	0.64	0.36
Roman Catholic	1208	563	0.68	0.32
Communist	298	7	0.98	0.02
Unknown	2	86	0.02	0.98
Neutral	11	55	0.17	0.83
Jewish	39	0	1.0	0.0
Christian	5	17	0.23	0.77
Center-Right	0	16	0.0	1.0
Anti-Revolutionaries	0	14	0.0	1.0
Patriotic	0	5	0.0	1.0
Conservative	0	3	0.0	1.0
Independent	1	2	0.33	0.67
Anti-Semitic	0	1	0.0	1.0
National Socialist	1	0	1.0	0.0

Table 7: *Distribution of articles between geographical scope and pillar.*

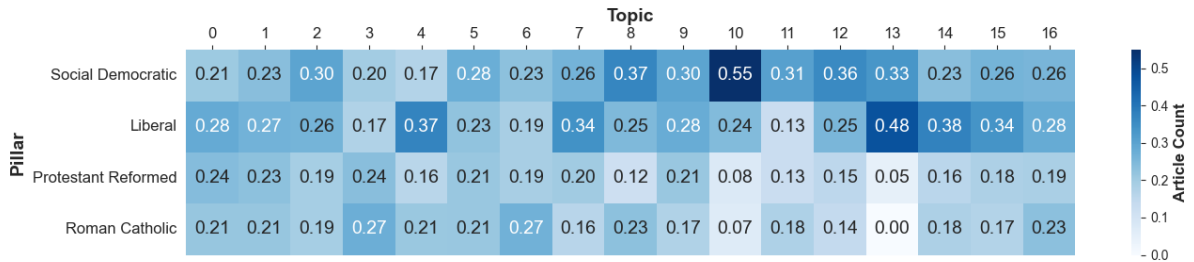


Figure 11: *Heatmap of the normalized proportion of articles per pillar for every topic.*

Starting with the one article published in the self-proclaimed anti-Semitic newspaper, ‘De Misthoorn’. However, this specific article is not connected to their ideology directly. The article, published in February 1942, is about the working conditions of people separating waste. This came up in search results, due to the used feminine pads (*damesverband*). The topic assigned to this article was Topic 12 (Daily Life & Intimacy). This article fits with the daily life aspects of this topic, with it being the conditions workers face during their work.

The other pillar with only one article was published by *Het Nationale Dagblad: voor het Nederlandse volk*, in 1940, a National Socialist newspaper. In this case, their ideological opinion did come across in the article. It was categorized under Topic 16 (Society, Sexuality & Puberty), which was fitting due to the article being about public health, such as teaching children to eat healthy and the continued research of bacteria. However, the article continued towards the well-being of individuals, and how genetic illnesses and conditions could be prevented by a marriage ban or sterilization.

The three conservative articles were published in 1952 by the same local newspaper, *Dagblad voor Noord-Limburg*, and they covered three different topics: 0 (Abortion Policy & Public Health), 4 (Women’s

Bone Health), and 3 (Religion & Cultural Views). The first two articles were about a Sister who obtained her PhD in medicine, with her dissertation on amenorrhea (absence of menstruation). The article with Topic 3 was about the mental health of the Catholic community.

The independent articles covered the topics 6, 0, published by a local newspaper, and 10, published by a national newspaper, respectively, published in 1950, 1948, and 1941.

An interesting thing is that the eight articles mentioned by now are all uploaded in the forties and early fifties. For over half, the reason is that the newspaper that published the articles was only issued around that time, and was discontinued shortly after, or was merged with other newspapers.

Social Democratic is the most frequent pillar. Figure 12 shows the distribution of the topics over the twentieth century. It shows that the peak of Topic 10 (medical Questions & Advice) that we saw in Figure 5 in the 30s is mostly represented by the social democratic. Further examination into this showed that it was mostly one specific newspaper (Het Volk/Het Vrije Volk), a social democratic newspaper that had a rubric on medical questions.

Topic 15 is the second most frequent for this pillar, following the overall number of articles. However, topic 0, which is also quite frequent, but compared to the full dataset, is much less.

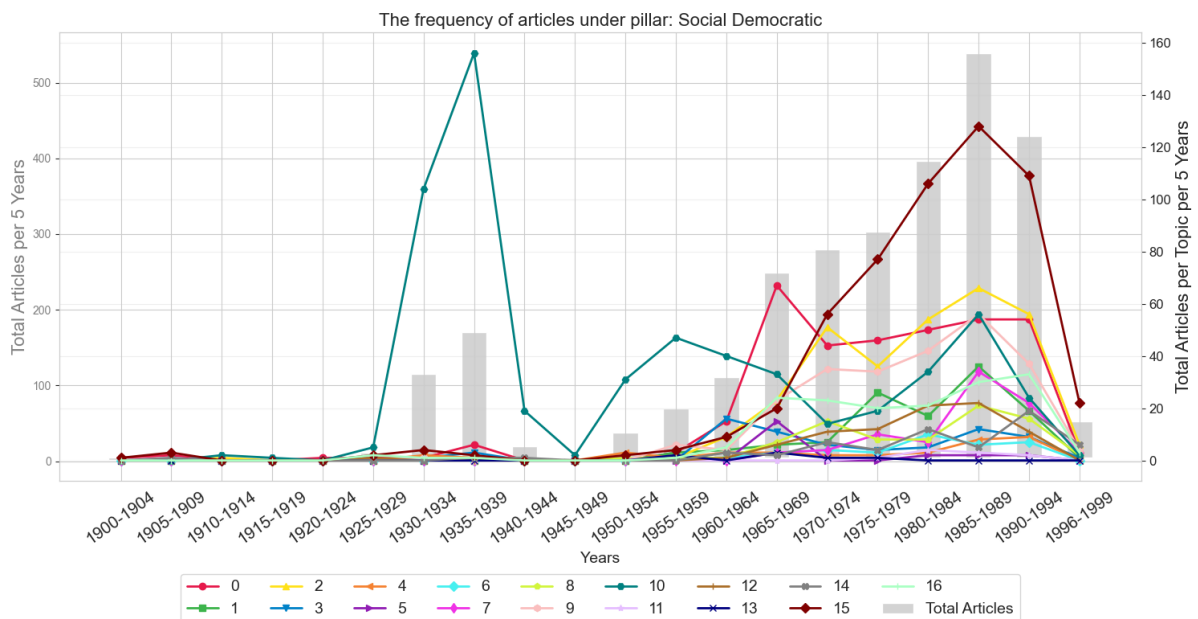


Figure 12: *The frequency of articles published by Social Democratic newspapers every 5 years for each topic.*

The distribution of the topics of articles published by Liberal newspapers is visualized in Figure 13, Topic 15 is the most frequent, following the overall trend of published newspapers. Topic 0 about Abortion Policy is the second most frequent topic. The liberal newspapers follow the rise of overall published articles with Topic 0; however, the social democrats do not.

Figure 14 shows the distribution of the articles published by Protestant newspapers. They were not very present during the early half of the twentieth century, and only started to publish stories more frequently containing menstrual-related topics in the early sixties. The early sixties covered mostly

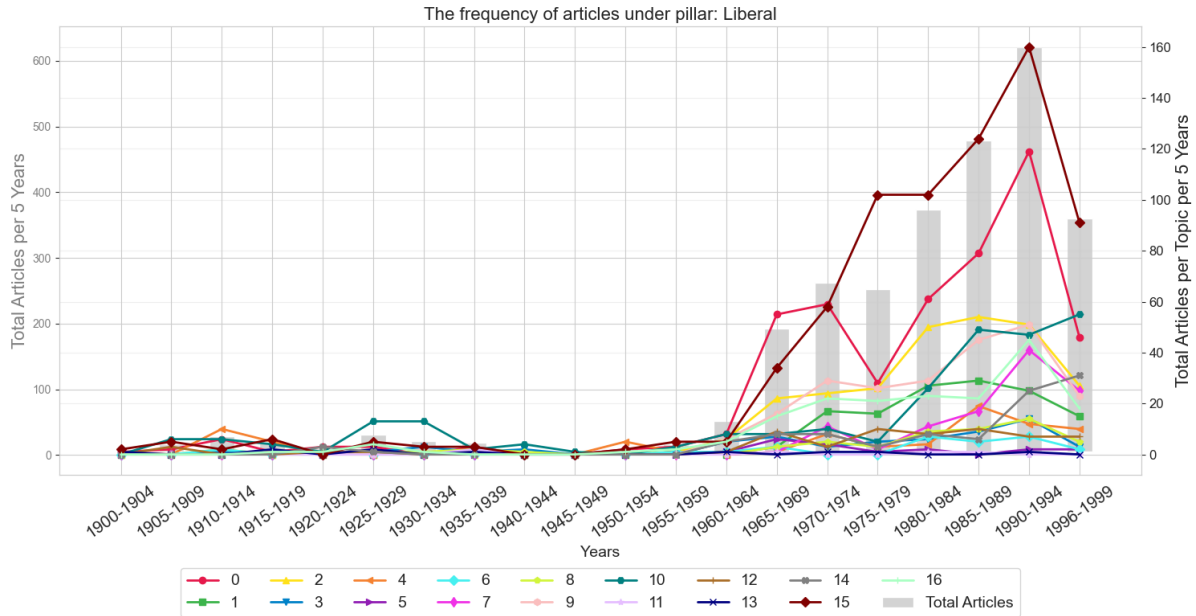


Figure 13: *The frequency of articles published by Liberal newspapers every 5 years for each topic.*

abortion policies (Topic 0) and Religions & Cultural views (Topic 3). This second topic is often how menstrual-related keywords are mentioned in the context of religion, about how the local churches and communities look at menstruation, abortion, or contraceptives. These topics start to rise with the overall number of articles, with a small dip of Topic 0 in the late seventies, which is probably due to the overall dip in total articles (as can be seen in Figure 5).

In comparison to the Protestant articles, Catholic articles are covered even more by Topic 3 (Religion & Cultural Views). After reading a random set of articles, one specific theme emerged: people were scared that the contraceptive pill could lead to being permanently sterilized. And newspapers often called in doctors with knowledge on it to say that the pill would not lead to sterilization. This opinion of sterilization came from Pope Pius XII, who said that the stopping of ovulation would lead to sterilization. Topic 3 would become less covered after the late sixties, when the topics of abortion and contraceptives became more popular. Because these topics have overlapped, it probably only means that the religious views became less pronounced after the sixties, which fits in with the pillarization (*verzuiling*) becoming less pronounced after the sixties.

From the total dataset of 9712 articles, 76% were published by national newspapers, while 24% came from local or regional newspapers. This distribution is consistent across nearly all topics, showing that there is a relatively uniform representation of national versus local press across the themes.

Table 8 presents the number and proportion of articles per topic per scope. These results show that the geographical division remains largely stable, regardless of subject matter. The only outlier is Topic 11 (Skin, Hormones & Pop Culture), where the majority (59%) was published by local newspapers. After a closer examination of the result, it became clear that most of these were from regional newspapers, those that publish across a province rather than a single city or town. These kinds of papers can serve as hybrids, combining the depth of national coverage with local accessibility, which may explain the higher

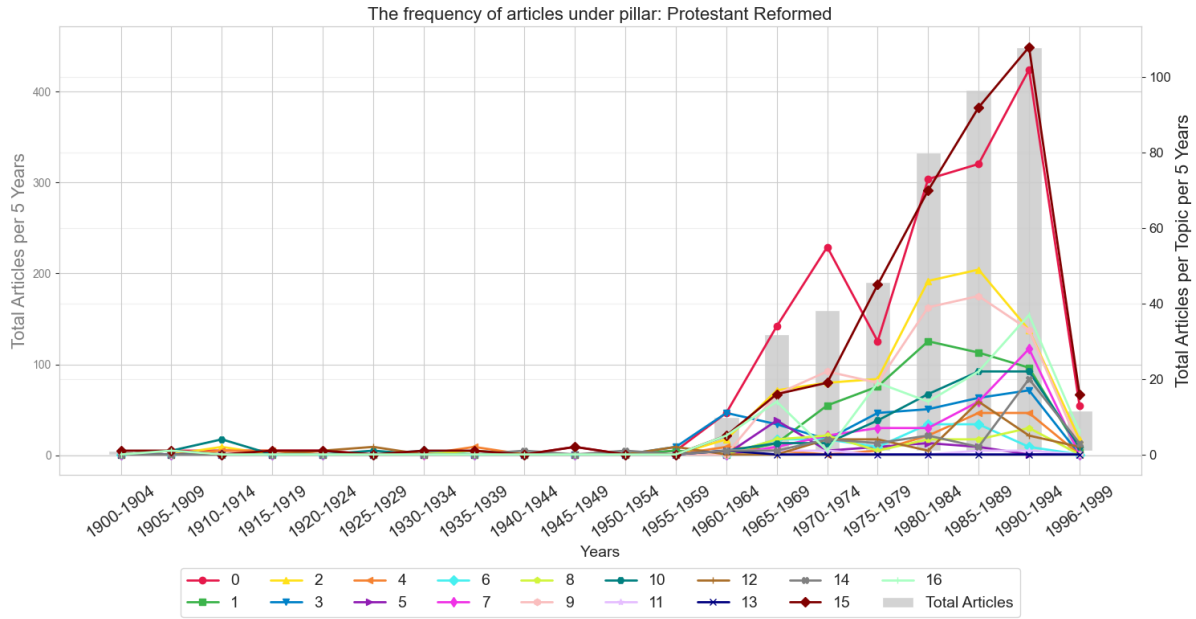


Figure 14: *The frequency of articles published by Protestant Reformed newspapers every 5 years for each topic.*

representation in this topic. For example, an article on what acne is and how to handle it.

National newspapers typically published a broader range of content, such as medical breakthroughs, international news, scientific research, and national laws and regulations, such as abortion policies or contraceptive regulation. In contrast, local newspapers focus more on community news and practical advice columns. However, the results show that national newspapers were still dominant across most topics.

7.2.3 Sentiment Analysis

To explore whether political or ideological orientation influences how menstruation was framed in Dutch newspapers, the sentiment-labeled articles were grouped by pillar. Table 9 shows both the absolute and proportional distribution of sentiment (neutral, positive, and negative) across each pillar.

For the four dominant pillars (Social Democratic, Liberal, Protestant Reformed, and Roman Catholic), the proportional sentiment distribution closely mirrors the overall trend observed in the dataset. This consistency suggests that, regardless of ideological orientation, the majority of articles treated menstruation in a factual, emotionally neutral tone.

The Liberal pillar stands out slightly for its relatively higher proportion of negative sentiment (5%), which may be linked to broader media coverage on gender representation and critiques found under Topic 15 (Media & Female Representation). In contrast, Social Democratic newspapers showed a marginally higher share of positive sentiment, possibly reflecting more progressive views on reproductive health and women’s rights.

Among the smaller pillars, more variation is visible; however, caution is warranted due to small sample sizes. For instance, Christian newspapers show the most extreme skew toward neutrality (91%),

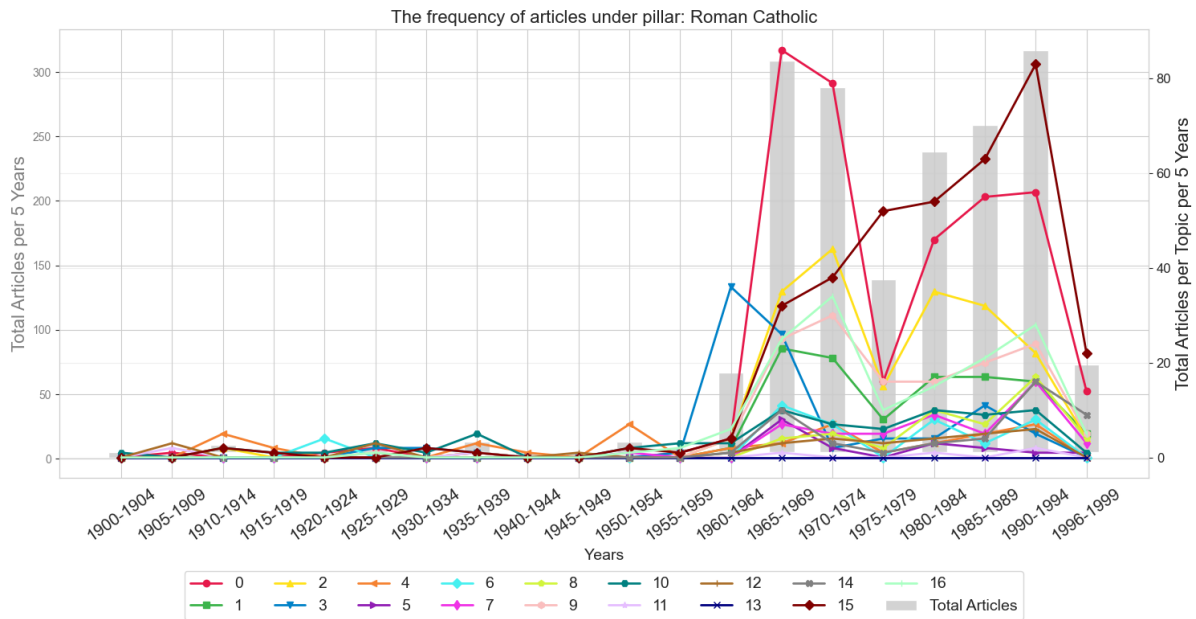


Figure 15: *The frequency of articles published by Roman Catholic newspapers every 5 years for each topic.*

while Center-Right newspapers have the highest share of negative sentiment (13%). However, these percentages are based on a limited number of articles, and true interpretation of these results should be done with care.

Articles from newspapers classified as having an unknown or neutral pillar also tended to show predominantly neutral sentiment. This aligns with expectations, as many of these are local publications without a clear or declared ideological affiliation.

Comparing the sentiment with the geographical scope did not reveal any major insights. 80% of the negative articles were published by national newspapers, compared to the 75-77% for neutral and positive articles.

In conclusion, sentiment distribution across political and ideological lines remained relatively stable, with neutrality dominating across all groups. While minor differences exist, particularly among smaller or more ideologically distinct pillars, these are generally not substantial enough to suggest strong bias, especially given the limitations of sample size and historical variance in tone and reporting style.

7.3 Discussion

This section demonstrated that both ideology and geography shaped how Dutch newspapers approached menstruation in the twentieth century, but in different magnitudes. National newspapers were the primary platform for discussions, representing all pillars. Their broader reach and focus on policy, science, and national debates made them the main medium for topics such as abortion, contraceptives, and medical advancements. From an ideological perspective, different pillars led different conversations around menstruation. Liberal and Social Democratic newspapers were leading, with a focus on medical questions, female representation, and gender. Protestant and Roman Catholic outlets, especially in the

Geographical Scope		Absolute		Proportional	
Topic	Name	National	Local	National	Local
15	Media & Female Representation	1680	418	0.8	0.2
0	Abortion Policy & Public Health	1285	403	0.76	0.24
10	Medical Questions & Advice	828	245	0.77	0.23
2	Contraceptives & Side Effects	733	278	0.73	0.27
9	Fertility & Reproductive Biology	608	201	0.75	0.25
16	Society, Sexuality & Puberty	474	167	0.74	0.26
1	Menopause & Physical Discomfort	378	130	0.74	0.26
3	Religion & Cultural Views	230	114	0.67	0.33
7	Estrogen & Cancer Risk	268	47	0.85	0.15
14	Economics & Industry	214	46	0.82	0.18
12	Daily Life & Intimacy	162	74	0.69	0.31
4	Women’s Bone Health	150	68	0.69	0.31
8	Hormones & Biology	175	43	0.8	0.2
6	Menstrual Blood & Hygiene	107	46	0.7	0.3
5	Women’s Mental & Hormonal Health	53	27	0.66	0.34
11	Skin, Hormones & Pop Culture	16	23	0.41	0.59
13	Cultural Perspectives on Menstruation	14	7	0.67	0.33

Table 8: *Distribution of the geographical scope per topic.*

1960s, engaged with menstruation through moral and religious views, often in reaction to the rise of contraceptives and abortion policies.

The overall tone of the newspapers remained neutral, and the ideological affiliation showed only subtle influences in the differences in sentiment. In addition to this, the geographical scope seemed to have no influence on the topics discussed or the sentiments expressed in the newspaper articles about menstruation.

Importantly, these results likely show the decline of pillarization in the second half of the century, which may have enabled newspapers to be more open to discussion of menstruation, particularly in national media.

8 Discussion

This thesis examined how menstruation was discussed in twentieth-century Dutch newspapers. By analyzing 9712 articles from the Delpher archive through a combination of topic modeling and sentiment analysis, five sub-questions were addressed, each contributing to a broader understanding of how menstruation was talked about in Dutch newspapers over time. Seventeen distinct topics were identified using Latent Dirichlet Allocation (LDA), with overall sentiment classified as predominantly neutral (76%), followed by positive (20%) and negative (4%).

Sentiment	Absolute			Proportional		
Pillar	Neutral	Positive	Negative	Neutral	Positive	Negative
Social Democratic	2062	622	116	0.74	0.22	0.04
Liberal	2113	532	132	0.76	0.19	0.05
Protestant Reformed	1360	383	58	0.76	0.21	0.03
Roman Catholic	1404	303	64	0.79	0.17	0.04
Communist	239	58	8	0.78	0.19	0.03
Unknown	75	12	1	0.85	0.14	0.01
Neutral	54	10	2	0.82	0.15	0.03
Jewish	30	8	1	0.77	0.21	0.03
Christian	20	2	0	0.91	0.09	0
Center-Right	12	2	2	0.75	0.13	0.13
Anti-Revolutionaries	12	2	0	0.86	0.14	0
Patriotic	4	1	0	0.8	0.2	0
Conservative	3	0	0	1	0	0
Independent	3	0	0	1	0	0
Anti-Semitic	1	0	0	1	0	0
National Socialist	1	0	0	1	0	0

Table 9: *Distribution of articles by sentiment per pillar.*

8.1 Interpretations

The first sub-question examined how the quality of OCR (Optical Character Recognition) influenced the retrieval of the articles from the Delpher archive. The results showed that while the majority of the corpus was legible and useful, the accuracy of OCR varied significantly depending on publication year, newspaper layout, font, and paper quality. Earlier documents, especially those published before the 1940s, were more prone to OCR errors. This occasionally led to missed documents or corrupted keyword matches. These issues highlight that the gathered corpus is not a comprehensive representation of the historical data and that OCR quality can impact the scope of historical research.

The second sub-question addressed how semantic ambiguity in the keywords *ongesteld*, *maandstonde*, *maandverband*, and *tampon* could be resolved in historical newspaper analysis. These words had second meanings unrelated to menstruation, depending on the time period or context. For example, *ongesteld* once meant generally ‘unwell’, and *tampon* could refer to a buffer or plug in French-language contexts. These ambiguities were handled by applying a text classification model, trained on data labeled by other menstrual-related keywords. While this approach helped improve the quality of the dataset, it also emphasized the need for domain-specific disambiguation tools when working with historical texts.

The third sub-question explored the topics related to menstruation discussed in Dutch newspapers and how these changed over time. Seventeen distinct topics were found with Latent Dirichlet Allocation (LDA), ranging from *Abortion Policy* and *Medical Questions* to *Media Representation* and *Fertility*.

The temporal distribution showed a clear rise in menstruation-related articles from the 1960s onward, reflecting broader changes in society, such as the sexual revolution, the rise of second-wave feminism, and advancements in reproductive health technologies. Several peaks were observed in the topics that

could be tied to historical events. For instance, Topic 0 (Abortion Policy) surged during the political debates of the 1970s and 80s, while Topic 2 (Contraceptives) appeared more frequently after the pill was introduced in 1962. The rise of articles under Topic 9 (Fertility) corresponded with advancements in reproductive technology. A unique trend was a rise of Topic 10 (Medical Questions & Advice) during the 1930s, with smaller peaks in the 60s and 80s, due to recurring medical Q&A columns in Social Democratic newspapers.

The fourth sub-question covered the sentiment attached to the articles and how this varied across topics and time. Using a fine-tuned version of the RobBERT Dutch language model, it was found that 76% of the 9721 articles expressed a neutral tone, 20% were positive, and only 4% were negative. Even though menstruation can be an emotionally charged or stigmatized topic, the newspapers often framed it in a medical or informational tone.

When broken down by topic, some interesting differences appeared. Articles covering Topic 9 (Fertility & Reproductive Biology) had a notably higher proportion of positive sentiment compared to the other topics, likely due to the optimism around medical advances like IVF. In contrast, Topic 16 (Society, Sexuality & Puberty) also had a higher proportion of positive sentiments; however, for this, there was no clear explanation. Topic 14 (Economics & Industry) had proportionally the most neutral articles, reflecting their focus on business reporting and pharmaceutical stock prices. However, caution is needed when interpreting the negative class, as the model performed poorly during evaluation due to the small sample size and low F1-score.

The final sub-question investigated whether the topic and sentiments varied according to the ideological background (pillarization) and geographical scope (national vs. local newspapers). Social Democratic and Liberal newspapers published the most articles on menstruation, followed by Protestant and Roman Catholic papers. 98% of Liberal newspapers were National, and most Unknown and Neutral newspapers were Local, due to them often not writing from a specific ideological perspective. Topic patterns also varied by pillar. For example, *Medical Questions & Advice* (Topic 10) was especially prominent in Social Democratic papers, and *Religion & Cultural Views* (Topic 3) was mostly found in Roman Catholic newspapers, especially in the 1960s, when contraception and abortion became prominent issues. However, the proportion of national compared with local coverage remained fairly stable across topics, with around 75% of the articles coming from national outlets. Sentiment patterns across pillars were also relatively consistent. Liberal newspapers showed a slightly higher percentage of negative sentiment (5%). However, for smaller ideological groups, the sample sizes were too small to draw strong conclusions.

This study shows how menstruation was discussed in twentieth-century Dutch newspapers, and that menstruation is connected to broader societal discourse, such as medial, political, moral, and cultural discussion. It was rarely treated as a standalone issue; it was often mentioned in connection with debates around contraception, abortion, gender norms, and medical progress. The majority of neutral sentiment may reflect both journalistic style and societal discomfort around menstruation in public media. Understanding how menstruation was portrayed in newspapers over time can also inform current discussions on menstrual stigma, policy, and education.

8.2 Limitations

Several limitations affected this study. First, the OCR quality of historical newspapers introduced noise and likely led to missed or misclassified articles. Second, the ambiguity of Dutch keywords related to menstruation required additional selection, which may not have captured all relevant cases, or could have included false positives. This could have had an effect on the results of the topic modeling. Third, the LDA model, while sufficient for the topic modeling, is not optimized for long-term temporal trends, and better-suited methods, such as Correlated Topic Modeling, were not explored due to time and tooling constraints. The sentiment classifier, although fine-tuned on a labeled sample, was limited by class imbalance and by the mismatch in historical and modern training data. As a result, the negative sentiment predictions were interpreted with caution.

The visualization of the results was one of the challenging aspects of the project. This applied both to reporting the combinations of topic, sentiment, geographical scope, and pillar affiliation in tabular form, and to visualizing the temporal trends of the topic modeling and sentiment analysis. While the visualization successfully reports the full findings, however, in the graph of the temporal topic modeling results, the overlapping of the multiple lines occasionally makes the results difficult to interpret.

Lastly, the author's background in AI rather than history or gender studies may have shaped the selection and interpretation of sources. While the project was driven by personal interest and curiosity, some historical or sociocultural nuances may be underrepresented.

8.3 Recommendations

Future work could expand the fine-tuning of the sentiment model by including a larger, manually annotated training set, ideally labeled by human experts with knowledge of historical context. Exploring other types of sentiment analysis, such as Aspect Based Sentiment Analysis.

For topic modeling, experimenting with dynamic topic models or BERTopic could result in more temporally sensitive insights. Correlated Topic Modeling was reviewed in the background section of this thesis; however, it was not used due to a lack of resources and time. In future works, exploring this model could give more interesting results.

While the quality of the OCR is the responsibility of Delpher, looking into models that are specialized in finding and restoring the errors is an interesting path that could help many projects in the field of digital humanities.

On the historical side, further research could include qualitative close reading of selected articles per topic, sentiment, or pillar to validate the modeling results. Collaboration between computational and humanities scholars would help refine and deepen the analysis. Additionally, exploring newspapers from other time periods, or comparing Dutch topics and sentiments with other countries, would offer a more comparative perspective.

9 Conclusion

This thesis set out to answer the following research question: *What topics and sentiments about the menstrual cycle were expressed in Dutch newspapers in the twentieth century?*

The findings show that news articles did discuss menstruation, sometimes explicitly, but more often indirectly, and typically framed in neutral or factual language. Mentions of menstruation and related topics increased sharply in the 1960s, reflecting changing attitudes, technologies, and policies. Seventeen distinct topics were identified with Latent Dirichlet Allocation (LDA), with some showing clear historical spikes linked to real-world events like the introduction of the contraceptive pill or the legalization of abortion.

The sentiments of the articles were mostly neutral, which aligns with its journalistic and factual approach, but the number of negative sentiments was less than initially thought.

Interestingly, while geographical scope (national vs. local) had little influence on the topics or sentiments expressed by the articles, ideological orientation did show differences in both topic and sentiment expression. This shows that understanding historical media through the lens of social structure and ideology is important.

Due to time and computational constraints, certain aspects, like improving the classification of ambiguous articles or fine-tuning the sentiment analysis model, could not be further improved. Future research could include more linguistic research, more analysis of other media forms, or comparing it with different cultures.

In conclusion, this thesis contributes to a growing field of digital humanities research by showing how computational methods can show historical patterns in sensitive or stigmatized topics.

10 Acknowledgments

First, I would like to thank my thesis supervisor, Martha Larson, for her guidance, feedback, and encouragement throughout the process.

I am also grateful to my roommates, Marieke and Rick, for patiently listening to me talk about complex problems and being my rubber ducks during the difficult parts.

A special thanks to Delpher for providing access to their newspaper archive and API.

To everyone who supported me during the journey, thank you.

References

- [1] Nationale Bibliotheek. Tweede Wereldoorlog, July 2025.
- [2] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.
- [3] Mark Boukes, Bob van de Velde, Theo Araujo, and Rens Vliegthart and. What’s the Tone? Easy Doesn’t Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, 14(2):83–104, 2020.
- [4] Annick C. van Brouwershaven, Sophie H. Bolt, and Jeroen G. F. Jonkman. The ongoing and contentious coverage of abortion in a progressive context: a long-term cross-outlet assessment of dutch abortion news (2000–2022). *Culture, Health & Sexuality*, pages 1–18, 2024.
- [5] Natalie Brown, Laura J Forrest, Rebekah Williams, Jessica Piasecki, and Georgie Bruinvels. ‘everyone needs to be educated’: pupils’ voices on menstrual education. *Reproductive Health*, 21(1):121, 2024.
- [6] Kyoung Won Cho, Shine Young Kim, and Young Woon Woo. Analysis of Women’s Health Online News Articles Using Topic Modeling. *Osong Public Health and Research Perspectives*, 10(3):158–169, June 2019.
- [7] Joan C. Chrisler. Teaching Taboo Topics: Menstruation, Menopause, and the Psychology of Women. *Psychology of Women Quarterly*, 37(1):128–132, 2013.
- [8] Ilana Cohen. Menstruation and Religion: Developing a Critical Menstrual Studies Approach. In *The Palgrave Handbook of Critical Menstruation Studies*. Springer Nature, 2020.
- [9] Vered Daitch, Adi Turjeman, Itamar Poran, Noam Tau, Irit Ayalon-Dangur, Jeries Nashashibi, Dafna Yahav, Mical Paul, and Leonard Leibovici. Underrepresentation of women in randomized controlled trials: a systematic review and meta-analysis. *Trials*, 23(1):1038, 2022.
- [10] Shelby H. Davies, Miriam D. Langer, Ari Klein, Graciela Gonzalez-Hernandez, and Nadia Dowshen. Adolescent Perceptions of Menstruation on Twitter: Opportunities for Advocacy and Education. *Journal of Adolescent Health*, 71(1):94–104, March 2022.
- [11] De Bruyne, Luna and De Clercq, Orphée and Hoste, Veronique. Emotional RobBERT and insensitive BERTje : combining transformers and affect lexica for Dutch emotion detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (EACL 2021)*, pages 257–263. Association for Computational Linguistics, 2021.
- [12] Pieter Delobelle, Thomas Winters, and Bettina Berendt. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, November 2020. Association for Computational Linguistics.

- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [14] Vimal Dixit, Kamlesh Dutta, and Pardeep Singh. Word Sense Disambiguation and Its Approaches. In *CPUH-Research Journal*, volume 1, pages 54–58, 2015.
- [15] Zulfadzli Drus and Haliyana Khalid. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science*, 161:707–714, 2019. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
- [16] Déborah Fusaro. Menstruation in news media: The impact of media discourses on the menstrual taboo in France. Master’s thesis, Lund University, 2016.
- [17] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the History of Ideas Using Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’08*, page 363–371, USA, 2008. Association for Computational Linguistics.
- [18] Arafat Hossain, Md. Karimuzzaman, Md. Moyazzem Hossain, and Azizur Rahman. Text Mining and Sentiment Analysis of Newspaper Headlines. *Information*, 12(10), 2021.
- [19] Anurita Jalan, Himansha Baweja, Mehar Bhandari, San Kahmei, and Aruna Grover. A Sociological Study of the Stigma and Silences around Menstruation. *Vantage: Journal of Thematic Analysis*, 1(1):47–65, April 2020.
- [20] Ingrid Johnston-Robledo, Jessica Barnack-Tavlaris, and Stephanie Wares. “Kiss Your Period Good-Bye”: Menstrual Suppression in the Popular Press. *Sex Roles*, 54:353–360, 11 2006.
- [21] Manju Kaundal and Bhopesh Thakur. A Dialogue on Menstrual Taboo. *Indian Journal of Community Health*, 26(2):192–195, June 2014.
- [22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, 2020.
- [23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.
- [24] Emily A. Marshall. Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics*, 41(6):701–724, 2013. Topic Models and the Cultural Sciences.
- [25] Celeste Nieuwendijk. Menstruatie in de media: een rode lijn tussen taboe en consumptie. Master’s thesis, Universiteit Utrecht, 2019.

- [26] NVSH. Nederlandse Vereniging voor Seksuele Hervorming (NVSH), 2025.
- [27] National Library of the Netherlands. OCR, June 2025.
- [28] Sonia Ponzo, Aidan Wickham, Ryan Bamford, Tara Radovic, Liudmila Zhaunova, Kimberly Peven, Anna Klepchukova, and Jennifer L Payne. Menstrual cycle-associated symptoms and workplace productivity in US employees: A cross-sectional survey of users of the Flo mobile phone app. *DIGITAL HEALTH*, 8:20552076221145852, 2022.
- [29] Annabelle Pronk. Social Politics of the Menopause. Master’s thesis, Leiden, 2018.
- [30] Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.
- [31] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020.
- [32] M. Sivakami, Anna Maria van Eijk, Harshad Thakur, Narendra Kakade, Chetan Patil, Sharayu Shinde, Nikita Surani, Ashley Bauman, Garazi Zulaika, Yusuf Kabir, Arun Dobhal, Prathiba Singh, Bharathy Tahiliani, Linda Mason, Kelly Alexander, Mamita Thakkar, Kayla Laserson, and Penelope Phillips-Howard. Effect of menstruation on girls and their schooling, and facilitators of menstrual hygiene management in schools: surveys in government schools in three states in India, 2015. *Journal of Global Health*, 9:13, 12 2018.
- [33] Teketo Tegegne and Mitike Sisay. Menstrual hygiene management and school absenteeism among female adolescent students in Northeast Ethiopia. *BMC public health*, 14(1):14, 10 2014.
- [34] Juha Teperi and Matti Rimpelä. Menstrual Pain, Health and Behaviour in Girls. *Social science & medicine*, 29(2):163–169, 1989.
- [35] Maria Kathryn Tomlinson. Moody and monstrous menstruators: the Semiotics of the menstrual meme on social media. *Social Semiotics*, 31(3):421–439, 6 2021.
- [36] Thijs van Dooremalen and Justus Uitermark. The framing of 9/11 in American, French, and Dutch national newspapers (2001–2015): An inductive approach to studying events. *International Sociology*, 36(3):464–488, 2021.
- [37] Renske Mirjam van Lonkhuijzen, Franshelis Katerinee Garcia, and Annemarie Wagemakers. The Stigma Surrounding Menstruation: Attitudes and Practices Regarding Menstruation and Sexual Activity During Menstruation. *Women’s Reproductive Health*, 10(3):364–384, September 2023.
- [38] Margot Vancauwenbergh and Karlien Franco. Women, blood, and dangerous things: socio-cultural variation in the conceptualization of menstruation. *Language and Cognition*, 16(2):505–535, August 2023.

- [39] Aditi Vashisht, Rambha Pathak, Rashmi Agarwalla, Bilkish Patavegar, and Meely Panda. School absenteeism during menstruation amongst adolescent girls in Delhi, India. *Journal of Family & Community Medicine*, 25(3):163–168, 09 2018.
- [40] Daniel Walker, William B. Lund, and Eric K. Ringger. Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. In Hang Li and Lluís Màrquez, editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 240–250, Cambridge, MA, October 2010. Association for Computational Linguistics.
- [41] Daniel Walker, Eric Ringger, and Kevin Seppi. Evaluating supervised topic models in the presence of OCR errors. In Richard Zanibbi and Bertrand Coüasnon, editors, *Document Recognition and Retrieval XX*, volume 8658, page 865812. International Society for Optics and Photonics, SPIE, 2013.
- [42] Liesbet van Zoonen. Rethinking women and the news. *European Journal of Communication*, 3(1):35–53, 1988.
- [43] Lisa Zuidema, Eveline M van Luik, Manna A Alma, Jaklien C Leemans, Marlies Y Bongers, and Peggy M A J Geomini. Informational needs related to menstrual literacy among Dutch women: a focus group study. *BMC women's health*, 25:158, 04 2025.
- [44] Esmée Zwiers. De opmars van de pil werd gehinderd door morele bezwaren. *Nederlands Interdisciplinair Demografisch Instituut*, 39(7):1–4, July 2023.

11 Appendix

A Topic Modeling Results

0		1		2	
abortus	abortion	overgang	menopause	bijwerking	side effect
minister	minister	lichamelijk	physical	anticonceptie	contraception
volksgezondheid	public health	pijn	pain	Haspels	Haspels
commissie	committee	psychisch	psychological	markt	market
advies	advice	verschijnsel	phenomenon	werking	function
gisteren	yesterday	gevoel	feeling	spiraal	IUD
wet	law	verandering	change	condoom	condom
verbieden	ban	hoofdpijn	headache	innemen	take
betalen	pay	allerlei	all kinds	farmaceutisch	pharmaceutical
ziekenhuis	hospital	migraine	migraine	methode	method
3		4		5	
kerk	church	osteoporose	osteoporosis	syndroom	syndrome
katholiek	catholic	bot	bone	depressie	depression
eeuw	century	vitamine	vitamin	Defares	Defares
god	god	proefschrift	dissertation	wetenschappelijk	scientific
verschijnen	appear	kalk	calcium	artikel	article
paus	pope	anorexia	anorexia	bevalling	childbirth
cultuur	culture	pleister	band-aid	vaatziekten	vascular disease
beschouwen	consider	dik	fat	studie	study
schrijver	writer	heup	hip	hartinfarct	heart attack
onderwerp	subject	Enk	Enk	hormonaal	hormonal
6		7		8	
bloed	blood	oestrogenen	estrogens	eierstok	ovary
tampon	tampon	borstkanker	breast cancer	oestrogeen	estrogen
dier	animal	oestrogeen	estrogen	mannelijk	male
vlees	meat	overgang	menopause	produceren	produce
plant	plant	risico	risk	cel	cell
ijzer	iron	hart	heart	progesteron	progesterone
kruis	cross	kanker	cancer	dier	animal
bacterie	bacteria	effect	effect	hersenen	brain
slip	slip	oestrogeon	estrogen	hypofyse	pituitary
afscheiding	discharge	vaatziekte	vascular disease	bloed	blood

9		10		11	
eicel	egg cell	bloed	blood	acne	acne
baarmoeder	uterus	huid	skin	mannelijk	male
ovulatie	ovulation	operatie	surgery	hond	dog
methode	method	eten	eating	kat	cat
bevruchting	fertilization	pijn	pain	effect	effect
cyclus	cycle	antwoord	answer	groeneveld	groeneveld
eisprong	ovulation	aandoening	condition	speelfilm	feature film
zaadcel	sperm	borst	breast	alberda	alberda
bevrucht	fertilized	verschijnsel	symptom	anabole	anabolic
gynaecoloog	gynaecologist	advies	advice	huid	skin
12		13		14	
eten	eat	modern	modern	miljoen	million
bed	bed	traditioneel	traditional	gulden	guilders
beetje	little	bevolking	population	bedrijf	company
huis	house	menopause	menopause	markt	market
half	half	beleven	experience	product	product
stoppen	stop	Israël	Israel	stijgen	rise
praten	talk	stad	city	prijs	price
nacht	night	China	China	gemiddeld	average
thuis	home	Chinees	Chinese	VS	US
hoeven	have	positief	positive	miljard	billion
15		16			
verhaal	story	seksueel	sexual		
film	movie	ouder	parent		
ding	thing	jongen	boy		
allemaal	all	school	school		
stuk	piece	vader	father		
kijken	look	gezin	family		
foto	photo	huwelijk	marriage		
mooi	beautiful	relatie	relationship		
vol	full	dochter	daughter		
dame	lady	seksualiteit	sexuality		

Table 10: *The Dutch (left) and English (right) words for the topic modeling results.*

B Newspaper Pillars & Scope

B.1 Frequency Pillars per Topic

Pillar \ Topic	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Social Democratic	348	119	303	68	37	22	35	83	80	242	589	12	85	7	61	545	164
Liberal	479	135	263	60	81	18	29	108	54	226	261	5	60	10	98	709	181
Protestant Reformed	402	116	194	81	35	17	29	63	26	173	84	5	36	1	41	379	119
Roman Catholic	362	109	193	94	46	17	41	51	50	141	75	7	33	0	47	357	148
Communist	59	24	46	12	1	2	6	10	6	25	40	2	7	0	7	47	11
Unknown	12	3	6	4	9	3	4	0	1	1	8	4	10	1	6	15	1
Neutral	15	1	3	2	3	0	4	0	1	1	6	2	2	0	0	20	6
Jewish	4	0	0	10	1	0	0	0	0	0	2	0	1	1	0	13	7
Christian	1	0	0	4	4	0	1	0	0	0	5	1	1	0	0	5	0
Center-Right	2	1	0	4	0	1	0	0	0	0	1	0	0	1	0	4	2
Anti-Revolutionaries	2	0	2	4	0	0	0	0	0	0	0	1	0	0	0	4	1
Patriotic	0	0	1	0	0	0	3	0	0	0	1	0	0	0	0	0	0
Independent	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
Conservative	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
National Socialist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Anti-Semitic	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

Table 11: *Distribution of articles per topic across the pillars.*

B.2 Newspapers Metadata

Newspaper	Frequency	Percentage	Scope	Pillar
Algemeen Dagblad	996	10.26	National	Liberal
De Volkskrant	910	9.37	National	Roman Catholic
Het Parool	890	9.16	National	Social Democratic
Het Vrije Volk	818	8.42	National	Social Democratic
Trouw	791	8.14	National	Protestant Reformed
De Telegraaf	746	7.68	National	Liberal
NRC Handelsblad	725	7.46	National	Liberal
Nieuwsblad van het Noorden	667	6.87	Local	Social Democratic
Leeuwarder Courant	626	6.45	Local	Protestant Reformed
Limburgsch Dagblad	383	3.94	Local	Roman Catholic
Nederlands Dagblad	353	3.63	National	Protestant Reformed
De Waarheid	270	2.78	Local	Communist
De Tijd	224	2.31	National	Roman Catholic
Het Volk	218	2.24	National	Social Democratic
Algemeen Handelsblad	144	1.48	National	Liberal
Het Rotterdamsch Parool	92	0.95	Local	Social Democratic
Tubantia	82	0.84	Local	Roman Catholic
De Nieuwe Limburger	67	0.69	Local	Roman Catholic
Het Vaderland	47	0.48	National	Liberal
Utrechts Volksblad	46	0.47	Local	Social Democratic
Nieuwe Rotterdamsche Courant	40	0.41	National	Liberal
Friese Koerier	38	0.39	Local	Liberal
Zaans Volkblad	37	0.38	Local	Social Democratic
Nieuw Israelietisch Weekblad	36	0.37	National	Jewish
Het Volksdagblad	35	0.36	National	Communist
Nieuwe Winterswijksche Courant	27	0.28	Local	Neutral
Nieuwe Haarlemsche Courant	23	0.24	Local	Roman Catholic
De Tijd De Maasbode	23	0.24	Local	Roman Catholic
De Maasbode	23	0.24	Local	Roman Catholic
De Gooi- en Eemlander	16	0.16	Local	Center-Right
Utrechtsch Provinciaal en Stedelijk Dagblad	14	0.14	Local	Neutral
Nieuwe Provinciale Groninger Courant	14	0.14	Local	Anti-Revolutionaries

Arnhemsche Courant	13	0.13	Local	Liberal
Leekster Courant	12	0.12	Local	Unknown
Nieuwe Tilburgsche Courant	11	0.11	Local	Roman Catholic
Nieuwe Vlaardingsche Courant	10	0.1	Local	Unknown
Winschoter Courant	10	0.1	Local	Social Democratic
De Courant	10	0.1	National	Neutral
Bredasche Courant	8	0.08	Local	Protestant Reformed
De Nieuwe Courant	8	0.08	National	Liberal
De Amsterdammer	8	0.08	Local	Christian
Het Nieuws van den Dag	8	0.08	Local	Liberal
Voorwaarts	7	0.07	National	Social Democratic
Utrechtsche Courant	7	0.07	Local	Neutral
De Tribune	7	0.07	National	Social Democratic
De Avondpost	6	0.06	Local	Unknown
De Vrije Socialist	6	0.06	National	Social Democratic
Ons Noorden	6	0.06	Local	Unknown
Nieuwe Eindhovense Krant	6	0.06	Local	Roman Catholic
Nieuwe Utrechtsche Courant	5	0.05	Local	Christian
De Standaard	5	0.05	National	Protestant Reformed
Zutphensche Courant	5	0.05	Local	Patriotic
De Nederlander	4	0.04	National	Christian
De Zaanlander	4	0.04	Local	Liberal
Land en Volk	4	0.04	National	Liberal
Overijsselsch Dagblad	4	0.04	Local	Unknown
Provinciale Drentsche en Asser Courant	4	0.04	Local	Unknown
Provinciale Overijsselsche en Zwolsche Courant	4	0.04	Local	Unknown
Nieuwsblad van Friesland	4	0.04	Local	Unknown
De Sumatra Post	3	0.03	Local	Unknown
Nieuwe Apeldoornsche Courant	3	0.03	Local	Unknown
Dagblad voor Noord-Limburg	3	0.03	Local	Conservative
De Noord-Ooster	3	0.03	Local	Protestant Reformed
Haagsche Courant	3	0.03	Local	Neutral
Friesch Dagblad	3	0.03	Local	Protestant Reformed
Nieuwsblad voor de Hoeksche Waard en IJselmonde	3	0.03	Local	Unknown
Gereformeerd Gezinsblad	3	0.03	National	Protestant Reformed

Emmer Courant	3	0.03	Local	Liberal
Tilburgsche Courant	3	0.03	Local	Roman Catholic
Delftsche Courant	3	0.03	Local	Neutral
Het Centrum	3	0.03	National	Roman Catholic
Centraal blad voor Israëlieten in Nederland	3	0.03	National	Jewish
Het Binnenhof	2	0.02	Local	Roman Catholic
Franeker Courant	2	0.02	Local	Protestant Reformed
Twentsche Courant	2	0.02	Local	Roman Catholic
De Heerenveensche Koerier	2	0.02	Local	Independent
De Banier	2	0.02	National	Protestant Reformed
Nieuw Utrechtsch Dagblad	2	0.02	Local	Social Democratic
Nieuwe Hoornsche Courant	2	0.02	Local	Unknown
De Graafschapper	2	0.02	Local	Christian
Nieuwsblad van het Zuiden	2	0.02	Local	Unknown
Westlandsche Courant	2	0.02	Local	Unknown
Limburger Koerier	2	0.02	Local	Roman Catholic
Dagblad van Noord-Brabant	2	0.02	Local	Roman Catholic
De Nieuwe Aaltensche Courant	2	0.02	Local	Christian
Nieuwe Groninger Courant	2	0.02	Local	Unknown
De Stichtsche Courant	2	0.02	Local	Unknown
Hoornsche Courant	2	0.02	Local	Protestant Reformed
Leeuwarder Nieuwsblad	2	0.02	Local	Protestant Reformed
Nijmeegsch Dagblad	1	0.01	Local	Unknown
Maassluische Courant	1	0.01	Local	Unknown
Dagblad van Zuidholland en 's- Gravenhage	1	0.01	Local	Unknown
Nieuwe Hengeloosche Courant	1	0.01	Local	Unknown
Deventer Dagblad	1	0.01	Local	Unknown
Opregte Steenwijker Courant	1	0.01	Local	Unknown
Venloosche Courant	1	0.01	Local	Unknown
De Zuid Limburger	1	0.01	Local	Roman Catholic
Aaltensche Courant	1	0.01	Local	Unknown
De Noord-Brabanter	1	0.01	Local	Roman Catholic
Rotterdamsch Nieuwsblad	1	0.01	Local	Unknown
Opregte Haarlemsche Courant	1	0.01	Local	Unknown
t Nieuws voor Kampen	1	0.01	Local	Unknown
Groninger Courant	1	0.01	Local	Unknown

Provinciale Geldersche en Ni- jmeegsche courant	1	0.01	Local	Protestant Reformed
Nieuwe Harlinger Courant	1	0.01	Local	Unknown
De Morgen	1	0.01	National	Unknown
Mooi Limburg	1	0.01	Local	Neutral
Eindhovenensch Dagblad	1	0.01	Local	Liberal
Leidsch Dagblad	1	0.01	Local	Roman Catholic
Het Nationale Dagblad	1	0.01	National	National Socialist
Het Huisgezin	1	0.01	National	Unknown
Groninger Dagblad	1	0.01	Local	Unknown
Nieuw Nederland	1	0.01	National	Independent
De Misthoorn	1	0.01	Local	Anti-Semitic
De Courant Het Nieuws van den Dag	1	0.01	National	Neutral
Provinciale Noordbrabantsche en 's Hertogenbossche courant	1	0.01	Local	Unknown
Harlinger Courant	1	0.01	Local	Unknown
Zutphensch Dagblad	1	0.01	Local	Unknown
De Zuid-Willemswaard	1	0.01	Local	Roman Catholic
Dagblad De Amsterdammer	1	0.01	National	Christian
De Amstelbode	1	0.01	National	Roman Catholic

Table 12: *The complete list of newspapers and corresponding metadata.*