Bachelor thesis
Computing Science

Radboud University

Measuring Evolution of Cookie Dialogues

Author:
Violeta Sizonenko
s1024157

First supervisor:
Dr. Hugo (H.L) Jonker
hugo.jonker@ru.nl

Second supervisor:
Dr. Christine Utz
e-christine.utz@ru.nl

October 16, 2024

Contents

Chapter 1

Introduction

In recent years, cookie pop-ups have become a familiar part of the web browsing experience, informing users about data storage and sharing practices. However, these dialogues were not always mandatory, which led to inconsistencies and potential misuse. The ePrivacy Directive in 2002 marked the first significant effort to protect user privacy, but it was the introduction of the General Data Protection Regulation (GDPR) in 2016 that set a new standard for data protection in the European Union. By 2018, the enforcement of GDPR with strict penalties for non-compliance prompted widespread changes across the digital landscape.

Despite these advances, there remains a critical challenge: assessing the real impact of these regulations on the evolution of cookie dialogues across different regions and languages. This study seeks to address this issue by developing a methodology to track and analyze the changes in cookie dialogues over time. The aim is to evaluate how data protection regulations have influenced the adoption and adaptation of these dialogues, providing insights into their effectiveness in enhancing user privacy. To explore this, the central research question is:

How can we evaluate the impact of data protection laws on the evolution of cookie dialogues?

Following the GDPR enactment, France's Data Protection Authority (CNIL) implemented additional regulations in 2019, further strengthening these protections. Our focus will be on tracing key legislative developments in France related to cookie dialogue compliance and examining how they have influenced the evolution of cookie dialogues on French websites.

To answer the main question, we are going to investigate the following sub-question:

How can we leverage a web archive to create a timeline of cookie dialogue changes over a period of time for a specific case?

Throughout our case study, we aim to use the WayBack Machine alongside machine learning model to track and analyze changes in cookie dialogues across French websites. Our goal is to develop a methodology that can be applied to multiple languages within the European Union domain, helping to contribute in researching this area of web privacy.

Contributions. In response to the challenges of assessing the evolution of online privacy practices, this thesis proposes a structured approach to tracking the emergence and adaptation of cookie dialogues over time. By combining the archival capabilities of the WayBack Machine with the XLM-RoBERTa model for multilingual classification, the thesis provides insights into how GDPR and CNIL regulations shaped compliance behavior in the digital space.

The contribution lies not only in creating a dataset that tracks cookie dialogue adoption but also in its potential applications for policymakers, researchers, and privacy advocates. This data can be used to assess the effectiveness of privacy regulations, track compliance patterns, and even identify periods of heightened regulatory enforcement, such as when major fines were imposed. Additionally, the research introduces a scalable, automated approach to studying regulatory impacts, making it adaptable to future studies across different countries, languages, and timeframes.

Moreover, the methodology's flexibility allows it to be tailored to examine other web-based phenomena beyond cookie dialogues, contributing to broader research areas like privacy, security, and web analytics. The findings also serve as a useful baseline for developing more advanced tools to further refine cookie dialogue classification and improve accuracy in detecting regulatory shifts.

Ethical Considerations. This thesis places ethical considerations at the forefront of studying how web services adapt to privacy regulations guidelines. By Adhering to the Menlo Report's ethical framework [1], we ensure responsible research conduct.

We abstain from gathering personally identifiable or sensitive data during crawling and filter out illegal or unethical websites from the analysis. Our use of the WayBack machine aligns with digital heritage preservation, in accordance with its terms of service.

We uphold intellectual property rights and copyright regulations and implement responsible crawling practices to minimize disruption. These measures guarantee ethical research conduct, fostering trust and the responsible utilization of web crawling and the WayBack Machine.

[1] The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research.

Chapter 2

Background

This chapter provides the essential context needed to understand the evolution of cookie dialogues in response to Data Protection Regulations (DPRs). It covers key topics such as the regulatory framework, the technical tools used for data collection and analysis, and the relevant concepts of cookies and consent mechanisms. A detailed timeline of privacy regulation changes can be seen in Appendix A.

2.1 Data Protection Regulations (DPRs)

Data Protection Regulations (DPRs) are legal frameworks designed to protect individuals' personal data and ensure privacy. They establish guidelines for how personal data should be collected, processed, and stored, often imposing strict requirements on organizations to ensure compliance.

ePrivacy Directive (ePD). The ePrivacy Directive (ePD), also known as the "Cookie Directive," was enacted by the European Union in 2002 and later amended in 2009 [1]. It complements the GDPR by specifically addressing privacy and electronic communications. The ePD requires websites to obtain informed consent from users before storing or accessing information on their devices, most notably through cookies. This directive laid the groundwork for cookie consent mechanisms, emphasizing the need for transparency and user control over personal data long before the GDPR expanded these requirements across all forms of data processing.

General Data Protection Regulation (GDPR). The General Data Protection Regulation (GDPR), enacted in May 2018 [2], is a comprehensive regulation by the European Union aimed at enhancing data protection and privacy for all individuals within the EU. It introduces stringent consent requirements, mandates data breach notifications, and grants individuals the

[1] ePD 2009.
[2] GDPR 679/16.

right to access and delete their data. The GDPR has significantly influenced how organizations handle personal data, including the implementation of cookie consent mechanisms.

Commission Nationale de l'Informatique et des Libertés (CNIL). CNIL is the French Data Protection Authority responsible for ensuring the enforcement of GDPR and other data protection laws in France. CNIL has issued specific guidelines[3] on cookie usage and consent, which have shaped how French websites implement cookie dialogues.

2.2 Cookie Banners and Dialogues

Cookie banners and dialogues are mechanisms implemented by websites to obtain user consent for cookie usage. Since the introduction of the GDPR, these tools have undergone significant evolution in both design and functionality, aimed at ensuring compliance while maintaining user engagement.

The GDPR introduced specific regulations concerning cookie consent mechanisms, placing a strong emphasis on explicit user consent and the transparency of information regarding the types of cookies used and their purposes. Cookie banners, typically displayed as pop-ups when a user first visits a website, serve as the initial point of interaction. These banners inform users about the website's cookie usage and may provide options for managing cookie preferences.

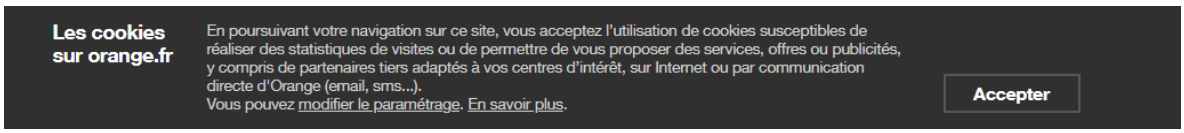


Figure 2.1: Example of a cookie banner before GDPR taken from https://www.orange.fr/ April 2018.

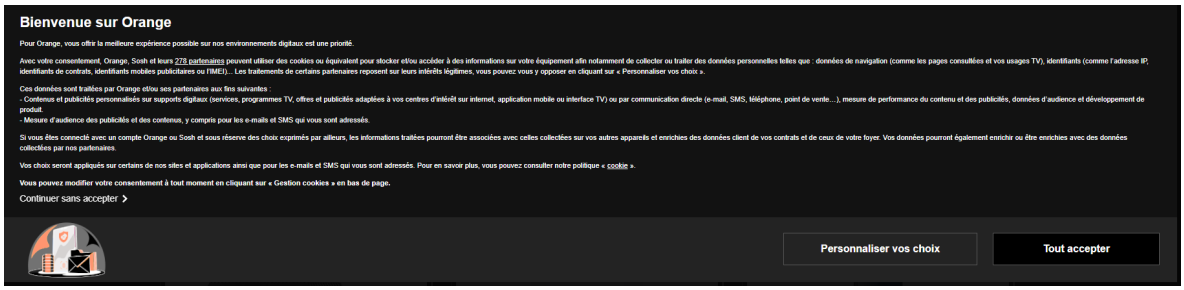


Figure 2.2: Example of a cookie dialogue after GDPR taken from https://www.orange.fr/.

Cookie dialogues go beyond basic notifications by offering users detailed information about the types of cookies used and the ability to accept or reject them. This interactive approach ensures that users are more informed and have greater control over their personal data, aligning with the regulatory requirements of the GDPR [4]. The shift from simple cookie banners to

[3] CNIL regulation.
[4] GDPR 679/16.

more comprehensive cookie dialogues represents an evolution driven by the GDPR's emphasis on user control and data transparency, an example for such change can be seen in Figures 2.1 and 2.2.

For a detailed overview of GDPR requirements related to cookies, please refer to Appendix A.

2.3 Tranco List

The Tranco list is a widely recognized and extensively used tool for web measurement and analysis. Similar to services like Alexa [5] and Majestic [6], Tranco provides a ranking of the top websites based on their popularity and visibility on the internet [LPVGT+19]. The list is compiled by collecting data from different sources, including web crawls and search engines, and ranks the top websites in descending order based on their traffic. While the ranking criteria may vary, they generally consider factors such as the number of unique visitors, incoming links, and domain authority. A Tranco list consists of a fixed number of websites referred to as the *Tranco Top N* websites. For a specific version of the list, the N value can vary, e.g. a Tranco list of the top 10,000 most popular websites. Tranco allows users to extract a copy of a list from a specific date, making it practical for historical analysis of changes in internet traffic over a specific period. Additionally, it allows customization to fit specific needs, such as filtering by country or region, adjusting list size, or including particular categories or domains, enabling targeted examination within specified parameters.

2.4 WayBack Machine

The WayBack Machine is a crucial tool for digital preservation, providing an extensive archive of the World Wide Web [7]. Its vast database is invaluable for researchers, historians, and anyone interested in studying the evolution of online content.

The WayBack Machine functions by using automated crawlers to visit and download web pages. Users can access this archived information by entering a URL into the search bar, which then displays a calendar view indicating the dates when the page was archived or using dedicated APIs.

- **Digital Archive**: Allows users to view historical versions of web pages, offering a comprehensive snapshot of the internet at various moments.
- **Extensive Database**: It has archived over 735 billion web pages [8], making it one of the largest digital archives globally.

[5] https://www.alexa.com
[6] https://majestic.com/reports/majestic-million
[7] http://web.archive.org
[8] The WayBack Machine general info.

2.5 Selenium

Selenium is an open-source framework designed to automate web browsing tasks [GGGMO20]. It enables developers and testers to simulate user interactions, conduct automated tests, and extract necessary data from web pages. Selenium facilitates precise interactions between code and web browsers, allowing for the automation of repetitive tasks and reduction of manual effort.

2.6 Web Crawling and Web Scraping

Web Crawling. Web crawling involves systematically browsing the internet to index web content. This process is essential in identifying and accessing relevant websites, as it navigates through pages and follows links to ensure thorough coverage.

Web Scraping. Following the identification of relevant websites through web crawling, web scraping is employed to extract specific data from these pages. This technique focuses on collecting precise information such as text, images, and links, enabling detailed analysis of the web content.

2.7 Bootstrapping

Bootstrap sampling is a statistical technique used to estimate the properties of an estimator, such as its variance, by sampling with a replacement from an observed dataset [ET93]. This method is particularly effective when dealing with small datasets or when the underlying distribution of the data is unknown. The process involves four key steps:

1. **Original Sample Selection**: Start with a single sample dataset, which forms the basis for creating the bootstrap samples. This dataset might consist of the observed data from your study, such as information on cookie dialogues.
2. **Resampling**: Generate multiple new samples, known as bootstrap samples, by randomly selecting data points from the original sample, with replacement. This means that the same data point can appear multiple times in a single bootstrap sample. Each bootstrap sample is the same size as the original dataset.
3. **Statistical Calculation**: For each bootstrap sample, calculate the desired statistic, such as the mean, median, or variance. This step results in a distribution of estimates, one for each bootstrap sample.
4. **Aggregation and Inference**: Analyze the distribution of the bootstrap statistics to draw inferences about the population. This can involve calculating the mean and confidence intervals for the statistic across all bootstrap samples, providing a robust estimate of the uncertainty or variability in the data.

In a research, bootstrap sampling can be used to assess the stability and reliability of observed data, ensuring that the results are consistent and representative, even with limited or varied data.

Chapter 3

Related Work

In the first part of this section, we will dive deeper into available research that aims to track cookie practice changes in response to data protection regulations. In the second part, we will talk about studies that used the WayBack Machine service for longitudinal researches.

3.1 Measuring the Impacts

The landscape of online privacy is intricate and continually evolving, with data regulations such as the GDPR aiming to enhance user control over personal information. However, assessing the real-world impact of these regulations on privacy practices presents significant challenges. This section reviews studies that delve into various methodologies and approaches for measuring and analyzing the effectiveness of privacy regulations, particularly focusing on consent notices and cookie consent mechanisms.

A study by Utz et al. examines the user interface of consent notices, a relatively unexplored area in GDPR compliance research [UDF+19]. Researchers aimed to understand common properties of consent notices by analyzing a dataset compiled from over 6,000 unique domains. This process involved using a Selenium-based automated browser setup to capture screenshots and manually check for the presence of consent notices. While the manual, detailed examination provides valuable insights, it also underscores the need for automated and scalable methods, which aligns with our research objectives.

Exploring another dimension of GDPR compliance, Soe et al. investigate the use of dark patterns in cookie consent mechanisms across 300 online news outlets [SNGS20]. Through a manual review of Scandinavian and English-language news websites, the researchers identified unethical practices designed to manipulate user consent. This work underscores the importance of scrutinizing design choices in consent mechanisms, which aligns with our goal of automating the detection and classification of such patterns on a larger scale.

To analyze the GDPR impacts on browser cookies, a study by Dabrowski *et al.* has conducted a longitudinal study on collected cookies from Alexa Top 100,000 websites where they compared the cookie behaviour from 2016 to that nowadays to see the change in behaviour. The study reveals that around 49.3% of Alexa Top 1,000 websites only set cookies after the consent is

granted when facing an EU visitor. This number drops by half when observing the Alexa Top 100,000 websites. These findings raise an important issue on the trend of less popular websites in slower adaptation to the privacy regulations.

Furthermore, a research by Zhuo et al. investigates the impact of GDPR on internet interconnection by analyzing traffic patterns before and after the regulation's implementation [ZHCG21]. By examining internet traffic data, the study identifies shifts in data flows and changes in privacy practices across networks, pinpointing alterations in data routing and processing practices attributable to GDPR. A key limitation noted was the difficulty in attributing all observed changes directly to GDPR due to concurrent global shifts in data practices. This study is particularly relevant to our method as it attempts to identify changes in traffic caused by data protection enforcement, reflecting our goal of detecting shifts in cookie practices.

These studies collectively illustrate the complexity of measuring the impact of privacy regulations and the necessity of employing diverse methodologies to capture the full scope of their effects. By examining these works, we have gained valuable insights into the limitations and challenges faced in previous research. These insights have directly influenced the development of our method, ensuring it is multi-linguistic, scalable, and automated. This comprehensive approach aims to address the gaps identified in prior studies and enhance our ability to analyze the evolution of cookie dialogues over time.

3.2 Leveraging the WayBack

The WayBack Machine, as an extensive web archive, has been used in various studies to explore the impact of privacy regulations over time. While its application in privacy research is still emerging, several studies have demonstrated its utility and highlighted its limitations.

To illustrate the efficacy of the WayBack Machine in recovering web content, Kumar *et al.* examined the rate of loss for online citations. They analyzed URL citations from journals and conferences to determine their persistence on the web. Using the WayBack Machine, they recovered content from vanished citations, proving its effectiveness. However, they also noted that only half of the web pages were archived, indicating a need for improvement in the service's coverage [KKP15, KP15]. These findings underscore the relevance of the WayBack Machine in tracking historical web data, which is central to our goal of monitoring cookie dialogues.

Building upon this foundation, Hashmi *et al.* used the WayBack Machine to study the evolution of ads and tracking domains over time. By collecting data from selected websites between 2009 and 2017, they analyzed changes in blacklists. The study pointed out limitations due to the WayBack Machine's redirections and inconsistent archiving frequencies, which could result in missed data [HIK19]. This highlights a critical aspect we need to consider in our longitudinal analysis of cookie dialogues

Further extending the application of the WayBack Machine, a study of Hadi *et al.* examined the evolutionary behavior of bug reports. Researchers explored the history of bugs, comparing resolved and open bugs over a decade. They employed a machine learning algorithm to validate their findings, showcasing the WayBack Machine's application in tracking web elements over

time [JCNS+22]. This aligns with our objective of analyzing cookie dialogues' evolution.

Similarly, Degeling *et al.* focused on changes in privacy policies post-GDPR implementation, using the WayBack Machine and a crawler for a semi-automated analysis across different countries. They found that increased transparency could lead to a false sense of security and suggested a multi-lingual approach [DUL+19]. This recommendation is pertinent to our research, as our model is designed for multi-lingual data analysis.

Finally, Dausend *et al.* evaluated GDPR compliance by manually checking 466 websites for cookie notices using the WayBack Machine. They observed minimal impact on cookie compliance practices in Germany and the US, highlighting the varied responses to GDPR [Dau23]. This study emphasizes the need for comprehensive, automated methodologies to better understand compliance trends across regions.

In conclusion, these studies collectively highlight the WayBack Machine's potential and limitations in longitudinal web data analysis. They inform our approach to developing a robust methodology for analyzing cookie dialogues over time, ensuring we address the challenges identified in previous research.

Chapter 4

Methodology

In this chapter, we outline a flexible methodology designed to study the evolution of cookie practices. Researchers can define their specific goals for the investigation, apply this methodology to a selected list of websites, and extract the first screen of cookie dialogues and their content from these sites over a chosen period. This process enables direct analysis of changes in cookie practices, designed to address specific research questions. This approach ensures that the methodology can accommodate a variety of research objectives and adapt to different investigative contexts.

The methodology for our study is structured into three distinct phases, each designed to support the overarching goal of capturing and analyzing the evolutionary changes in cookie dialogues. Each phase focuses on specific tasks that contribute to the comprehensive collection, classification, and analysis of data.

1. **Historical Web Data**: The first phase involves obtaining a list of pages from the WayBack Machine archive to ensure a diverse and representative dataset. The aim here is to efficiently gather a substantial number of web pages, laying the ground for in-depth analysis.
2. **Collecting Web Elements**: Once the web pages are identified, the next step is to systematically collect web elements from the list of pages. We employ web scraping techniques to extract detailed information about cookie dialogues and consent buttons on a large scale.
3. **Classifying cookie dialogues and buttons**: In this phase, the collected cookie dialogues and their associated buttons are subjected to detailed classification. We utilize multi-lingual data classification using XLM-RoBERTa model to analyze both the textual and structural elements of the cookie dialogues.

Each phase is designed to build upon the previous one, creating a layered approach to data collection and analysis that enhances the reliability and depth of our findings. Figure 4.1 illustrates the general idea of our method. We give an example of possible research tasks to highlight the idea that our method can be used to investigate different aspects in cookie dialogues.

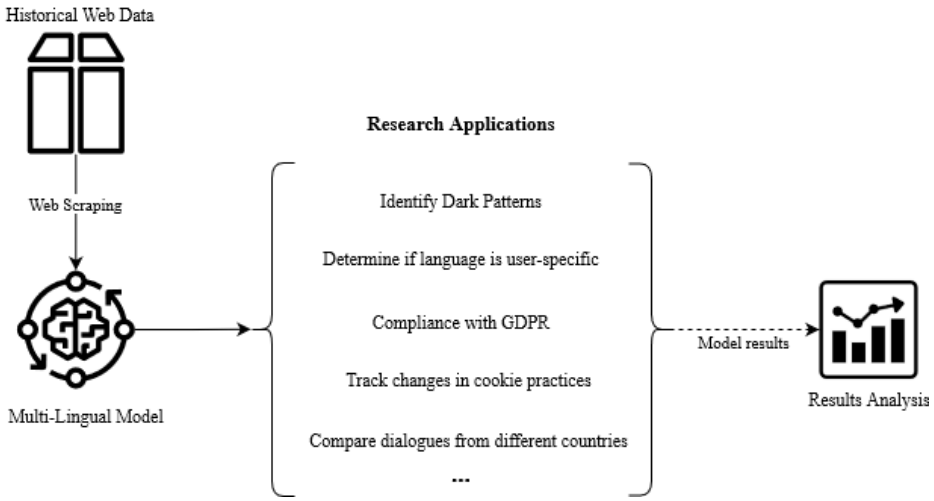


Figure 4.1: Method overview

4.1 Historical Web Data

Website Selection. In this research, we focus on analyzing the most popular websites with EU top-level domains (TLDs). While there are several services that offer metrics on website popularity, these rankings often vary, reflecting different measurement criteria and susceptibility to manipulation.

Services like Majestic SEO's Majestic Million and Alexa's Top Sites use distinct approaches to rank websites. Majestic Million, for example, ranks websites based on the number of subnets hosting a page and is updated daily, primarily focusing on backlink analysis. On the other hand, Alexa emphasizes traffic volume and audience demographics to determine its rankings. However, research by Le Pochat et al. [LPVGT+19] highlights the potential for manipulation in these popular rankings, demonstrating the extent to which services like Alexa, Cisco, Majestic, and Umbrella can be influenced. In response, these researchers introduced Tranco, a new ranking service that consolidates lists from Alexa, Cisco, Majestic, and Umbrella to create a more robust and less manipulable index.
Tranco enhances the validity, consistency, and verifiability of website rankings by filtering out unavailable and malicious domains. This approach results in a minimal daily change of at most 0.6%, providing a stable and reliable foundation for data collection. Given its resilience against manipulation and its comprehensive aggregation methodology, Tranco's ranking is chosen for our study to ensure the robustness and reliability of the data we collect.

Web Archive Selection. Selecting an appropriate tool for retrieving historical web data is essential for effectively studying the evolution of web content over time. The chosen tool must offer detailed chronological records, encompass a broad spectrum of websites, and ensure accurate capture of web content.

Archive.today [1] allows for manual, on-demand archiving of specific pages. Although this is useful for focused data collection, it does not provide systematic, periodic data collection necessary for broad temporal analyses. In contrast, search engine caches from providers like Google [2] and Bing [3] offer rapid access to recent web page snapshots. However, these caches do not maintain a long-term historical archive, making them unsuitable for longitudinal studies. Common Crawl [4] captures a wide selection of internet data monthly, which is useful for identifying broad trends. Yet, the monthly interval is too broad for capturing the nuances required by studies that need to track the impacts of specific events or regulatory changes more frequently.
In comparison, the WayBack Machine delivers frequent archiving intervals and extensive historical coverage, aligning with the needs for detailed and precise chronological research. It supports a comprehensive analysis across a diverse set of web domains, providing both the breadth and depth required for thorough historical analysis.
Therefore, the WayBack Machine emerges as the most suitable choice. It fulfills the stringent requirements for detailed, accurate, and extensive analysis of historical web data. While no tool can guarantee complete data capture, the WayBack Machine's robust archiving capabilities significantly reduce the likelihood of missing critical data, making it invaluable for research that demands high reliability and comprehensive scope.

4.2 Multi-lingual Data Classification

To classify cookie dialogues from web pages in multiple languages, we consider various machine learning models known for their robust handling of language nuances.

Before evaluating specific models, it is essential to understand the foundational technology upon which many of them are built. In the study by Van Hofslot *et al.* a classificaiton model BERT (Bidirectional Encoder Representations from Transformers) and its variations have been used to evaluate cookie banner legal regulation violations, focusing on the language used [VHASGS22]. The study showed BERT and its variation LEGAL-BERT had the highest accuracy (70%-97%). BERT is a breakthrough in the field of machine learning, particularly in natural language processing (NLP) [DCLT19]. By training language models based on the entire set of words in a sentence or query (bidirectional), it allows the model to grasp context more effectively, significantly enhancing its ability to understand and generate human-like responses. Several models were assessed for their potential in providing accurate and efficient multi-lingual classification:

mBERT (Multilingual BERT): An extension of the original BERT model, mBERT is trained on Wikipedia data in 104 languages, making it capable of processing text in multiple languages. However, its reliance on Wikipedia as a training dataset may limit its applicability to diverse web vernaculars [PSG19].

DistilBERT: This model offers a streamlined version of BERT that maintains much of the original model's performance but at a reduced complexity and resource requirement. While

[1] https://archive.ph.
[2] https://www.google.com.
[3] https://www.bing.com.
[4] https://commoncrawl.org.

efficient, its simplified nature may lack depth in capturing complex nuances in cookie dialogue analysis [SDCW20].

T5 (Text-to-Text Transfer Transformer): Unlike BERT, which directly handles classification and question-answering tasks, T5 converts language tasks into a unified text-to-text format, such as translating text or summarizing information, which could risk losing nuances in complex legal or technical terminology when applied to a multi-lingual dataset [RSR+23].

XLM-RoBERTa: After evaluating various models, XLM-RoBERTa emerges as the optimal choice for our multi-lingual classification needs. This model is a refined iteration of the original BERT architecture, significantly enhanced to facilitate multilingual processing. Unlike BERT and its direct successor, RoBERTa, XLM-RoBERTa leverages the extensive Common Crawl web data, which encompasses content in over 100 languages [LOG+19]. This broad and varied dataset provides XLM-RoBERTa with a broad linguistic context, significantly reinforcing its capacity to grasp and interpret language variations across diverse cultural backgrounds.

The effectiveness of XLM-RoBERTa extends beyond its theoretical design. Empirical studies, such as those conducted by Conneau et al., have shown that XLM-RoBERTa surpasses various BERT variations, including mBERT, in multiple Natural Language Processing (NLP) tasks. It excels in text classification and sentiment analysis across different languages, demonstrating superior performance [CKG+20]. These capabilities render XLM-RoBERTa a valuable tool for our project, for a precise analysis of web content in numerous languages.

RoBERTa-LSTM Hybrid: In the exploration of effective models for multi-lingual classification, the hybrid model that combines RoBERTa with Long Short-Term Memory (LSTM) networks is worth discussing. This hybrid aims to merge RoBERTa's deep contextual understanding with the sequential data processing capabilities of an LSTM, potentially enhancing the model's ability to process and analyze lengthy and complex text sequences [TLAL22]. While theoretically advantageous, this combination introduces additional complexity into the model architecture. The integration of LSTM with RoBERTa's already robust framework might not yield equivalent benefits, especially when compared to more streamlined models. Therefore, we choose to use **XLM-RoBERTa** as it offers comprehensive training and effectiveness in multi-lingual environments that align with our research objectives.

4.3 Proof of Concept Implementation

To conclude the Methodology section, we present an overview of our Proof of Concept (PoC) implementation which can be seen on Figure 4.2. This diagram illustrates the practical application of the proposed method, showcasing the flow from data acquisition to analysis.

As depicted in Figure 4.2, the process begins with retrieving a list of popular websites using the Tranco list. These websites are then accessed through the WayBack Machine Archive via API calls to fetch historical web pages. The collected web pages undergo web scraping to extract relevant elements, specifically cookie dialogues and consent buttons. These elements are then analyzed using the XLM-RoBERTa classification model to accurately categorize cookie dialogues and buttons. The classified data is compiled into a collected list, which is subsequently analyzed to determine trends and patterns in cookie dialogue practices.

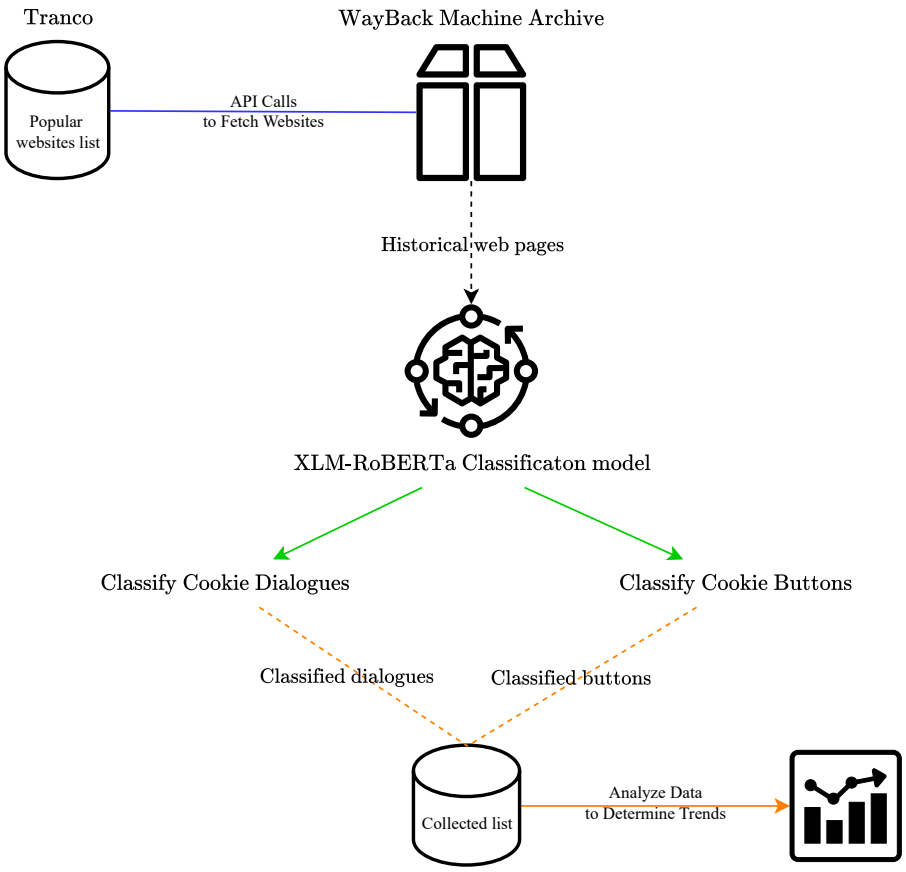


Figure 4.2: Proof of Concept Implementation Diagram

Automated Data Collection with Selenium and Firefox. For the automated collection of web elements, we use Selenium in conjunction with the Firefox browser. Selenium automates the browsing process, allowing us to systematically visit websites and interact with their cookie dialogues. This automation is essential for handling large datasets, ensuring that our study can scale effectively without manual intervention.

Cookie Dialogues and Banners. In our study, cookie dialogues and banners are treated as equivalent due to the ambiguity in regulatory guidelines from the GDPR and enforcement action from data protection authorities such as CNIL. These terms are often used interchangeably, and the lack of precise definitions in the legislation further complicates their differentiation as well as because the standards for either have evolved over time, both must be taken into account.

Chapter 5

Case study: evolution of French cookie dialogues

In this chapter, we provide a specific case for the proposed method in order to analyze the usability of it as well as to discuss the results. We start by explaining why this specific case study was chosen to evaluate the method. Then, we describe the experiment and glimpse over the implementation. In this section, we delve into the practical application of our proposed methodology, focusing on a specific case study: the evolution of cookie dialogues on French websites.

5.1 Case Study Set Up

To evaluate the effectiveness and feasibility of our proposed methodology, we intend to apply it to a specific case study framework. Our objective is to select a European country subject to the General Data Protection Regulation (GDPR) [1] with a strong Data Protection Authority (DPA) and a notable history of enforcing data protection laws. Additionally, we aim to demonstrate the effectiveness of our chosen machine learning model, XLM RoBERTa, by selecting a domain where the primary language is not English. Given these criteria, France, with its national DPA, the Commission Nationale de l'Informatique et des Libertés (CNIL) [2], emerges as an ideal candidate. This case study will specifically analyze the emergence and evolution of cookie dialogues and consent buttons over a period that correlates with GDPR and CNIL enforcement actions.

Focus on Popular Websites. To begin, we retrieve the Tranco Top 1 million most popular websites using the Tranco list [3] service generated on February 20, 2019. Although our original plan was to use data immediately following the GDPR compliance deadline, constraints led us to use the earliest available Tranco list. Despite the slight delay, this dataset remains representative of the Tranco Top 1 million popular websites at that time.

[1] GDPR 679/16.
[2] CNIL regulation.
[3] https://tranco-list.eu

We focus on these popular websites because they are under intense scrutiny from Data Protection Authorities (DPAs) due to their significant traffic volumes and the vast amounts of user data they process. These sites are more likely to be targeted by regulatory bodies for compliance checks and enforcement actions, making them prime candidates for studying the evolution of cookie practices. Their high visibility also means that any changes in their cookie policies are likely to influence broader industry trends, thus providing valuable insights into the effectiveness of GDPR and CNIL regulations. Furthermore, popular websites tend to update their content and user interfaces more frequently, ensuring that our analysis captures a dynamic and current perspective on compliance trends.

Date Range. To create a timeline of appearance of cookie dialogues, focusing on the periods influenced by GDPR and CNIL regulations, we have selected date range from April 2016 to April 2021. This aligns with key milestones in these regulatory frameworks, providing a comprehensive five-year period for analysis (see Appendix A for detailed dates).

- **GDPR Regulation (EU) 2016/679:** Date of issue: 27 April 2016; Adaptation date: 25 May 2018 [4].
- **CNIL:** Date of issue: 17 September 2020; Adaptation date: 31 March 2021 [5].

While we aim to trace the changes that GDPR has affected in the French services, it is important to take into consideration the state of the web privacy practices prior to the GDPR legislation for a deeper analysis. By examining the timeframe before its enforcement, we can capture the early compliance efforts and anticipatory changes, highlighting proactive versus reactive adjustments to the regulatory landscape. The end of range, on the other hand, is interesting because the potential for enforcement actions and penalties becomes a reality for organizations.

Other interesting dates that we believe have affected the introduction of cookie dialogues are the sanctions given by CNIL to Google and Amazon [6] [7] [8]. These fines, amounting to tens of millions of euros, were among the first major penalties for non-compliance with GDPR regulations in France, which likely prompted other service providers to accelerate their own compliance efforts. The importance of these dates lies in their potential to act as catalysts, pushing companies to swiftly update their websites to meet GDPR standards, thereby impacting the trend of cookie dialogue implementations.

Number of Websites. For this case study, we focus on analyzing 1,000 of the most popular websites in France: This selection provides a comprehensive and representative sample of the French internet landscape, enabling a thorough examination of cookie dialogue practices across a diverse range of sites. Choosing 1,000 websites is a balance between capturing a broad spectrum of data and managing the practical constraints of data processing. Given the complexity and time required for classifying cookie dialogues using the XLM-RoBERTa model, this sample size is both feasible and sufficient to yield meaningful insights while maintaining

[4] GDPR 679/16.
[5] CNIL regulation.
[6] Google 2019 fine by CNIL.
[7] Google 2020 fine by CNIL.
[8] Amazon 2020 fine by CNIL.

the depth and accuracy needed for the study.

5.2 Experiment

This part is split into three phases: Website collection, Web elements retrieval and Classification of the elements.

Website Collection. The website collection process involved executing scripts to filter the Tranco list according to our predefined criteria, focusing on selecting the Tranco Top 1,000 websites for French code top-level domain (ccTLD). The core of this phase was retrieving historical snapshots from the WayBack Machine. Using the WayBack Machine API, we accessed these snapshots on the service side, managing API data retrieval errors. It involved losing connection to the server or the absence of the record in the archive. The end of this phase was the organization and logging of successfully retrieved URLs into a structured JSON format, reflecting our specified date range and ccTLD, ensuring a comprehensive dataset for the subsequent classification process.

Web Elements Classification. After successfully generating the URL list, we extracted and classified relevant web elements from these websites. Implementing automated web scraping, we navigated through each website, parsing HTML content to target *iframe* and *div* elements — the most likely containers for cookie dialogues. We then used the XLM-RoBERTa machine-learning model to classify this extracted data to identify the presence of cookie dialogues and determine the types of buttons included. Finally, we documented the classification results and stored them in a JSON format corresponding to each website.

5.2.1 Experiment Implementation

Below, we are giving a general pseudo code for our implementation illustrating creation of list of URLs and Classification of Web Elements from these URLs. A more detailed specification of the WayBack API is given in Section 5.2.2.

Algorithm 1 Algorithm for creating a list of URLs

```
function FETCH_ARCHIVED_URLS(website, start_date, end_date, collapse_by, depth = 0)
    if depth > 1 then
        return ∅
    end if
    list_of_urls ← ∅
    reply ← WayBackAPI call to get URLs for website, start_date, end_date, collapse_by
    for each data_point in reply do
        try
            date ← extract date from data_point
            url ← extract URL from data_point
            list_of_urls.append((url, date))
        catch
            wait for 30 seconds
            return fetch_archived_urls(website, start_date, end_date, collapse_by, 1)
        end try
    end forreturn list_of_urls
end function
function GET_URLS()
    allWebsites ← Tranco API retrieve list for given date
    websites ← Filter allWebsites to end with .fr and to be of length of 1,000
    list_of_urls ← ∅
    for web in websites do
        list_of_urls[web] ← fetch_archived_urls(web, 2016-04, 2021-04, timestamp:6, 0)
    end for
    Save list_of_urls to json file
end function
```

Generating a list of URLs. The first part of our implementation process focuses on generating a comprehensive list of URLs from the Tranco list dated 2019-02-20.

1. **Tranco List Filtering**: From the extensive Tranco list, we filter out the Tranco Top 1,000 websites for our selected ccTLD - .fr.
2. **WayBack Machine API Usage**: Via *CDX waybackpy*, we access the WayBack Machine to find the closest historical record to our specified dates. The API attempts to retrieve a record for each date, moving to the next date with a specified step. If retrieval fails, the API continues attempts with the closest timestamp until successful.
3. **JSON File Creation**: For the ccTLD, we create a JSON file containing the successfully retrieved URLs within our date range, ensuring a structured and accessible dataset for analysis.

Algorithm 2 Help functions for for Classifying Web Elements

```
function CRAWL_URL(url_to_visit)
    try
        Visit url_to_visit using selenium
        Wait for wayback to redirect and website to load
        Get all text elements, buttons, iframes, divs
        Get all text and buttons elements from iframes and div with max depth 20 or time search 6 minutes
        Check for word "Cookie" in text and if found put it in the beginning of the list to check
        Check values for cookie dialogue
        if not cookie dialogue found then return "not found"
        end if
        Check for buttons for accept and decline return cookie dialogue, cookie buttons, "found"
    catch
        return "error"
    end try
end function
function COLLECT_WEBSITE_DATA(websites_to_crawl, dictionary_of_urls)
    for each web in websites_to_crawl do
        for each url_to_visit, date in dictionary_of_urls[web] do
            if dialogues_found ≤ 3 then
                Call crawl_url(url_to_visit)
                if found dialogue then
                    dialogues_found ← 0
                else
                    dialogues_found ← dialogues_found + 1
                end if
                if dialogues_found == 0 then
                    save a copy of results
                end if
                if found == "error" then
                    dialogues_found ← 0
                    Get value for web link to error
                end if
            else
                Break loop
            end if
        end for
        if not found a dialogue then
            Save "no dialogue found" for this website
        else
            Save all found dialogues
        end if
    end for
    return Save dialogues for all websites
end function
```

Algorithm 3 Algorithm for Classifying Web Elements

Part 2: Classifying Web Elements
dialuge_model, button_model ← TRAIN XLM-RoBERTa model on dataset for dialogue and button classification
Load list of website and their links to visit from json to urls_to_visit
Get websites_to_visit from urls_to_visit
Reverse links to visit in urls_to_visit
collect_website_data(websites_to_visit urls_to_visit)
STORE classification results in JSON file for URL

Classifying Web Elements. The second part involves classifying web elements using the XLM-RoBERTa machine learning model, focusing on identifying cookie dialogues and their corresponding buttons on the web pages:

1. **Model Training**: Initially, we train the XLM-RoBERTa model on a dataset comprising 650 entries for dialogue classification and 1,150 for button classification, covering all EU languages.
2. **Web Page Parsing**: With Selenium WebDriver in headless mode (browser windows are not visible), we process each URL from our JSON files, parsing the HTML to locate "iframe" and "div" elements that potentially contain cookie dialogues. We recursively go through these elements (because "iframe" might have "iframes" inside and so on) until we visit them all or we have reached the *depth 10* from the initial element.
3. **Elements Classification**: We extract text from these web elements, relying on the XLM-RoBERTa model to classify whether a cookie dialogue is present.
4. **Button Classification**: If the model determines the presence of cookie dialogue, it extracts the text content of potential buttons and classifies them.
5. **Misclassification Handling**: It performs one more iteration on the next date, and if it also contains a cookie dialogue, stop the iteration for this website. This way, we ensure it was not a misrecognition and that a cookie dialogue has indeed been added to the website at the spotted date. Otherwise, we continue analyzing with the next date.
6. **Output Compilation**: The final step involves compiling the classification results into a JSON file for each URL, documenting the dates and text of identified cookie dialogues and buttons.

The way the data range is examined is done in a reverse descending order - from April 2021 to April 2016. This serves our purpose better because of an assumption that most of web services do not tend to adapt to the regulations in the first half of the specified period, hence we can save some time by looking for a "disappearance" of a cookie dialogue and buttons rather than appearance. Additionally, we will omit potential dialogues with a text length shorter than 125 characters, as they may not comply with GDPR requirements simply because it is too short to fit the required content.

5.2.2 WayBack Implementation Specification

To discuss the choice of a day we are going to check for every month, we need to point out that from the WayBack side, the CDX Server's function:

```
WaybackMachineCDXServerAPI(url=website, user_agent=USER_AGENT,
    start_timestamp=start_date, end_timestamp=end_date, collapses=[
    collapse_by])
```

where the parameter *collapse_by* allows us to adjust the frequency of a check, it is limiting our choice of step size to three options: once a month, once every 10 days - 01, 10, 20, 30 of a month, and once a day. What is also important to note, is that upon requesting the URLs of a specified date, the API would return succeeding closest date. For example, if the specified timestamp is for every month, then the API will attempt to return the earliest available URL of the 1st of the month. This should not affect the results and the flow of the experiment as long as we keep it in mind.

To conduct the search, we will use the Selenium WebDriver package to access and manipulate web content, observing pages in headless mode. The WayBack server may refuse connections due to excessive requests, so we will use the *driver.wait()* function to avoid overloading the server.

5.2.3 Foreseen limitations

There are certain limitations that may affect the outcomes of the experiment. It is important to be aware of these limitations upon interpreting the results:

- Due to the time complexity of this task, the study focuses on 1,000 selected websites in France. It might lead to a not entirely accurate representation of internet landscape in France at that time, however, it should provide a confident insight.
- Multiple popular websites on the EU territory do not use EU ccTLD (e.g., bol.com) but still follow the GDPR guidance. At the end of this research, the presented data will not give an ultimate insight yet provide a general idea.
- The WayBack machine does not keep records of each website for any given day. For example, if a website has implemented a cookie dialogue on the 31st, but the WayBack only kept a copy of the 30th and the 2nd of the next month, the output will be the date of the following month. The results will contain a few weeks' error but will still be sufficient for our research question.
- The study's reliance on the earliest available Tranco list from 20-02-2019, instead of the initially intended 01-06-2018 list, presents a limitation regarding immediate post-GDPR most popular websites analysis and reproducibility. This gap might affect the ability to fully replicate or extend the study with data from the initial aftermath of the GDPR implementation.
- Due to the time feasibility, the crawler runs in headless mode - without displaying the interface and hence using less resources - which can potentially alter the web elements to the point a cookie dialogue might not show [KJK22].

Considering the limitations of the WayBack API, we expect a margin of flexibility of available URLs of approximately 10%, allowing us to account for missing URLs while still maintaining the total sample of 1,000 websites with 61 URLs each, one for each month in the time period.

5.3 Results

In this section, we present the results of our experiment creating timeline to track cookie dialogue evolution.

5.3.1 Results for Identifying Cookie Dialogues

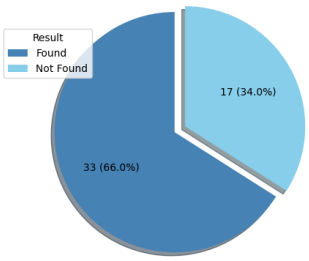
Presence of Cookie Dialogues

We observed the Tranco Top 50, Top 200, and Top 1,000 most popular websites in France over the five-year period to assess the presence of cookie dialogues. The findings are summarized in Figure 5.1, which categorizes the results into two groups: websites where a cookie dialogue was found and those where it was not.

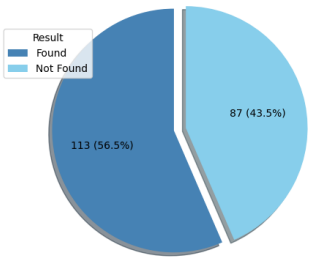
To be classified as a website with a cookie dialogue, the following criteria were applied:

- The length of the potential cookie notice text must exceed 125 characters.
- The content must be classified as a cookie notice by the XLM-RoBERTa model.

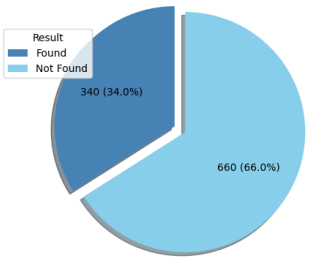
For the Tranco Top 50 websites, 66% were identified as having a cookie dialogue, while 34% did not. In the Top 200 websites, 56.5% were found to have a cookie dialogue, with 43.5% falling into the Not Found category. Among the Top 1,000 websites, 34% had a cookie dialogue, and 66% did not.



(a) Presence of cookie dialogues in Top 50 Websites.



(b) Presence of cookie dialogues in Top 200 Websites.



(c) Presence of cookie dialogues in Tranco Top 1,000 Websites.

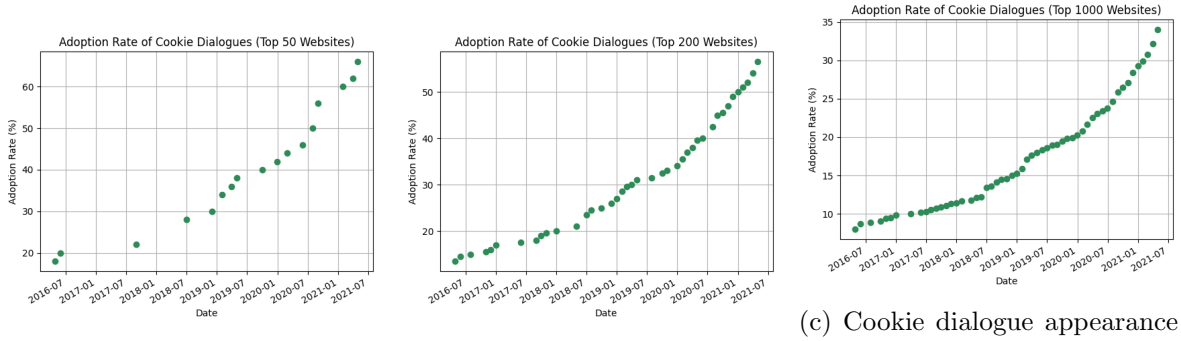
Figure 5.1: These pie charts illustrate percentages of identified and unidentified cookie dialogues for Top 50, 200 and 1,000 websites.

Adoption Rate

Figure 5.2 presents the adoption rates to cookie dialogue regulations in Top 50, Top 200 and Tranco Top 1,000 most popular websites for the period of five years.

In the Tranco Top 50 Websites, the adoption rate begins at around 20% in mid-2016 and steadily increases to over 60% by mid-2021, showing a consistent rise with significant increases post-2018. In the Top 200 Websites, starting at around 15% in mid-2016, the rate gradually

climbs to approximately 55% by mid-2021. For the Top 1,000 Websites the rate starts at around 10% in mid-2016 and reaches about 35% by mid-2021.



(a) Cookie dialogue appearance rate in Top 50 Websites.

(b) Cookie dialogue appearance rate in Top 200 Websites.

(c) Cookie dialogue appearance rate in Tranco Top 1,000 Websites.

Figure 5.2: These scatter plots illustrate the cookie dialogues implementation rate for Top 50, 200 and 1,000 websites.

Response Time

Figure 5.3 illustrates the response time to GDPR guidelines regarding cookie dialogues for Tranco Top 1,000 Websites. The cookie dialogues implemented before the five year period come up to 10%, the Period 3 accounts for the most common response time of 21.5% with the Period 4, 5 and 6 accounting for 18%, 17.6% and 19.7%.

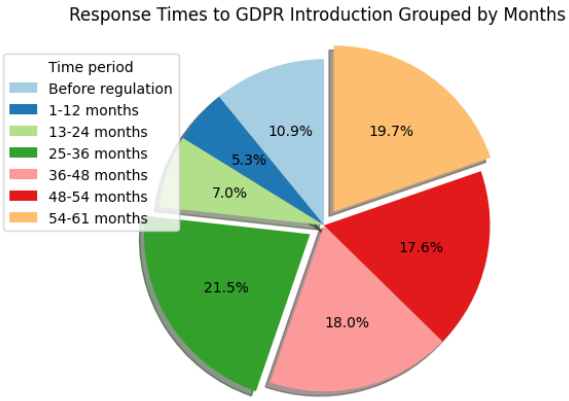


Figure 5.3: This pie chart demonstrates the response times of Tranco Top 1,000 Websites to GDPR guidelines.

Timeline for cookie dialogue evolution in France

We observed 1,000 most popular domains in France dated February 20, 2019. For the period of five years, each website has been checked monthly to spot the first appearance of a cookie dialogue in Tranco Top 50, Top 200 and Top 1,000 most popular websites. The data is

represented bimonthly per number of websites that introduced a cookie dialogue in that month. For a quick reference, GDPR and CNIL legislation were added as well as the biggest sanctions within that period. The results are summarised in Figure 5.4.

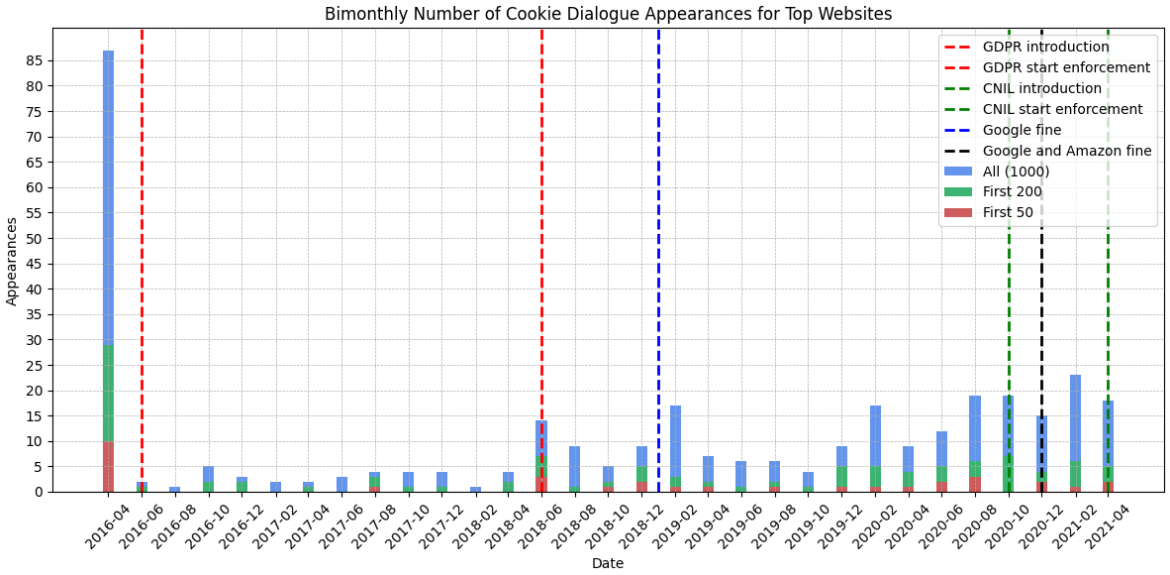
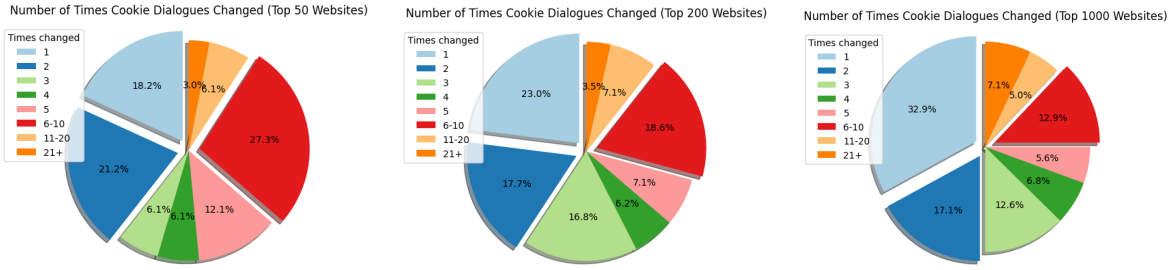


Figure 5.4: This stacked bar chart represents the timeline of cookie dialogues first appearances over the period of 5 years including the dates of legislative events in France.

5.3.2 Results for Tracking Changes in Cookie Dialogues

The pie charts in Figure 5.5 present the distribution of how many times the cookie dialogues changed on the Tranco Top 50, 200, and 1,000 websites with an identified cookie dialogue over the five-year period.

For the Top 50 websites, the majority, 27.3%, changed their cookie dialogues between 6 to 10 times, followed by 21.2% that changed it twice, and 18.2% changed it only once. The least frequent category was those changing their cookie dialogues more than 21 times, which accounted for 3.0%. In the Top 200 websites, 23.0% of them changed their cookie dialogues only once, and 18.6% of the websites had their cookie dialogues change between 6 to 10 times. Notably, 3.5% of websites changed their cookie dialogues over 21 times. In the Tranco Top 1,000 websites, 32.9% changed their cookie dialogues only once. A smaller percentage, 17.1%, changed their dialogues 2 times, while 12.9% of websites updated their dialogues between 6 to 10 times. Very few websites, only 5.0%, updated their cookie dialogues more than 21 times.



(a) Times cookie dialogues changed for Top 50 Websites.

(b) Times cookie dialogues changed for Top 200 Websites

(c) Times cookie dialogues changed for Top 1,000 Websites

Figure 5.5: These pie charts illustrate how many identified cookie dialogues changed over five year period for Tranco Top 50, 200 and 1,000 websites.

Figure 5.6 illustrates the changes in cookie dialogue lengths between the first and last occurrences, showing data only for websites that have changed their cookie dialogues at least twice. We filtered out values above 1,500 characters to enhance readability. For the Top 50 websites, the length difference varied up to 500 with the median of 200 characters, one outlier was detected; for the Top 200 websites, it varied up to 400 with the median of 150 characters, there were two outliers; and for the Top 1,000 websites, the range was up to 250 with the median of 100 characters while three outliers were present.

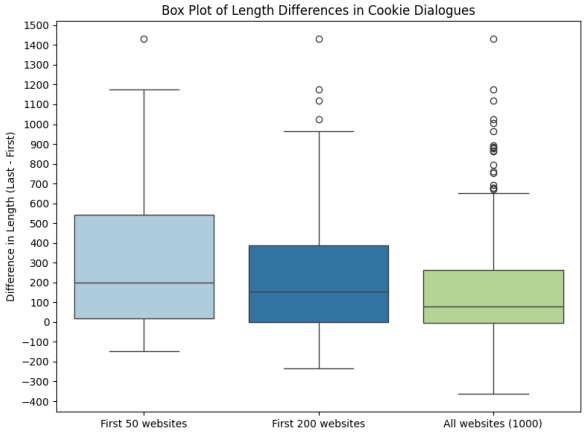


Figure 5.6: This box plot demonstrates the distribution of cookie dialogue length difference between first and last occurence for the list of Tranco Top 50, Top 200 and Top 1,000 websites.

5.4 Analysis

The results of our experiment of cookie dialogue appearance and evolution, as described in section 5.3, show the effects the GDPR and CNIL regulations had on the French websites. In this section, we will discuss the implications of these findings.

Analysis of Presence of Cookie Dialogues

The results reveal a clear trend in the adoption of cookie dialogues among popular websites in France over a five-year period. A significant majority of the Tranco Top 50 websites (66%) implemented cookie dialogues, likely due to their higher visibility and the increased scrutiny they face from regulatory bodies. However, as the scope expands to the Top 200 and Top 1,000 websites, the proportion of sites with detectable cookie dialogues decreases - from 56.5% to 34%. This suggests that smaller or less prominent websites may not prioritize GDPR and CNIL compliance as rigorously, highlighting a potential compliance gap.

Analysis of Adoption Rate

For the Tranco Top 50 websites, the adoption rate steadily increases from approximately 20% in mid-2016 to over 60% by mid-2021. This trend suggests a consistent adoption of cookie dialogues among the most popular sites, likely driven by the enforcement of GDPR and subsequent CNIL regulations. In the case of the Top 200 websites, the adoption rate shows a similar upward trajectory, but at a slower pace, reaching just over 50% by mid-2021. This indicates a broader, but slightly delayed, implementation of cookie dialogues across a wider range of popular websites. For the Top 1,000 websites, the adoption rate begins at around 10% in mid-2016 and gradually climbs to approximately 35% by mid-2021. This slower adoption rate among a larger pool of websites suggests that smaller or less trafficked sites may have been slower to implement cookie dialogues, possibly due to fewer resources or a perceived lower risk of regulatory punishment.

Analysis of Response Time

The most significant portion, 21.5%, represents websites that adjusted their cookie dialogues 13-24 months after the GDPR's introduction. The period just before the GDPR's enforcement saw 19.7% of websites implementing changes, highlighting a proactive approach by some services. However, 18.0% and 17.6% of websites only adjusted their practices 25-36 and 36-48 months post-GDPR, indicating a delayed compliance response. A smaller fraction, 7.0%, made changes in the earlier 1-12 months post-GDPR, while only 5.3% took action before the regulation came into effect, showcasing varying levels of urgency and preparedness across different websites. These findings suggest a significant delay in the overall compliance with GDPR, with nearly half of the websites taking more than two years to fully adapt to the regulatory requirements.

Analysis of Timeline

The data reveals several distinct periods of increased cookie dialogue appearances. Notably, there is a significant spike in the number of appearances around the time of the GDPR enforcement in May 2018. This spike is most prominent across all three groups of websites, indicating a substantial push towards compliance immediately following the regulation's enforcement. Subsequent spikes correspond to CNIL's introduction and enforcement of specific guidelines in late 2020 and early 2021. Particularly, the fines imposed on Google and Amazon

in December 2020 appear to have catalyzed another wave of cookie dialogue implementations, as seen by the sharp increase in appearances during this period. The pattern of spikes following major regulatory actions suggests a reactive approach among many websites, where compliance is largely driven by the introduction of new regulations or the threat of enforcement, rather than proactive measures. Overall, this graph highlights the strong influence of regulatory actions and enforcement on the adoption of cookie dialogues across popular websites in France.

Analysis of Number of Cookie Dialogue Changes

The analysis reveals that more popular websites, such as the Tranco Top 50, frequently updated their cookie dialogues, indicating a strong commitment to regulatory compliance. This behavior is likely driven by their high visibility and the higher chance of legislative prosecution. In contrast, as we move to the Top 200 and Top 1,000 websites, the frequency of updates decreases, suggesting that less popular websites may not have the same resources or incentives to stay as up-to-date. This indicates a more reactive approach to compliance among smaller websites and highlights the need for increased support and awareness to ensure broader adherence to privacy regulations.

Analysis of Cookie Dialogue Length Changes

We observe a consistent pattern where, on average, websites increased the length of their cookie dialogues over time. The interquartile range across all categories (Top 50, Top 200, and Top 1,000 websites) indicates that cookie dialogues grew in complexity, likely due to evolving legal requirements and the need for greater transparency. The presence of a few outliers suggests that some websites made significant changes, possibly as a reaction to specific compliance actions or penalties. The differences are more visible in the smaller datasets (Top 50 and Top 200), indicating that the most popular websites may have adapted faster or in more dramatic ways compared to the broader selection.

5.5 Validity

Despite the capabilities of the XLM-RoBERTa model and the archival data from the WayBack Machine, classification discrepancies such as false positives (instances where non-cookie dialogues were incorrectly classified as cookie dialogues) and false negatives (actual cookie dialogues that were missed by the classification) were encountered. We performed manual checks on the data selected by bootstrap sampling to confirm or refute its classifications. We have checked 400 websites with resampling from the list of one thousand. Through this verification, we were able to spot misclassifications, errors and to get an insight of how well our implementation performs.

5.5.1 Verifying WayBack API

The WayBack API successfully retrieved records for 980 out of the 1,000 websites. Upon reviewing the number of URLs generated per website, only 33% of the websites reached the desired count of 61 URLs as illustrated on Figure 5.7. Due to limitations of the WayBack API, several websites fell short of this target. To be more specific, we retrieved 51,991 URLs from the WayBack API out of the expected 61,000, which accounts for 85.23%. As we mentioned in the foreseen limitations in Section 5.2.3, we expected the margin of flexibility of 10%. However, the actual margin that we observed is 14.8%. Although it is exceeding the original margin, we still consider it reasonable given that we are investigating 1,000 French websites.

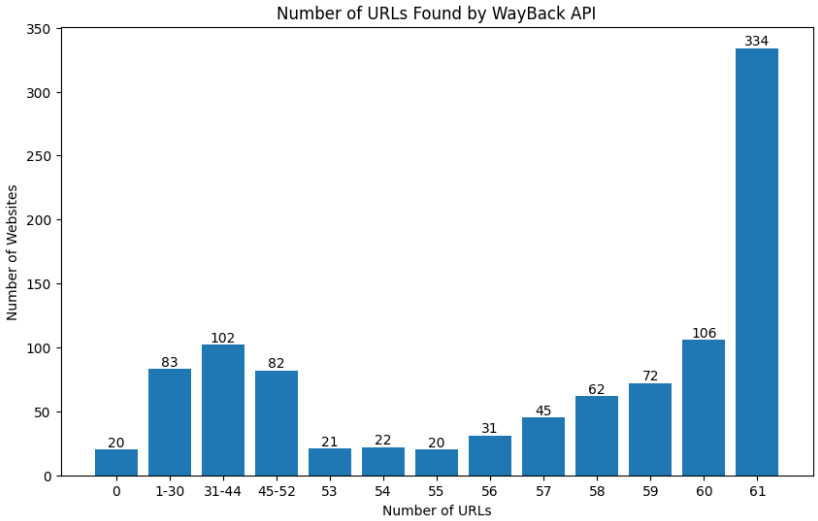


Figure 5.7: This graph show the distribution of the number of URLs obtained by WayBack API.

5.5.2 Verifying the Method with Bootstrapping

To evaluate the accuracy of the implemented methodology, bootstrapping was conducted across three lists of websites, with sample sizes of 40, 100, and 250. We assessed the accuracy of two key components: identifying the date of the first occurrence of a cookie dialogue and recognizing the content of the dialogue. Although the accuracy of detecting the first occurrence is dependent on correctly identifying the cookie dialogue content, it is critical to separate the results of our XLM-RoBERTa model from our crawler's performance.

For this purpose, we categorized data into two categories: **Dialogue Present** for websites that end up having a cookie dialogue on their page, and **No Dialogue** for websites with no cookie dialogue. While only **Dialogue Present** case was evaluated, the **No Dialogue** results showed near-perfect accuracy across all datasets, with almost no misclassification. As a result, we focus on presenting the accuracy of the model for **Dialogue Present** cases, where there is more variability and a need for deeper analysis. This allows us to highlight the areas where the system may require further improvement.

The Figure 5.8 represents the accuracy distribution for websites with cookie dialogue, broken down into **Correct Date** which stands for correctly identifying the date of first occurence of one and **Correct Text** that stand for correctly recognizing the content of a dialogue. The graph displays results from different sample sizes used in the bootstrapping analysis, allowing for a clear comparison across datasets.

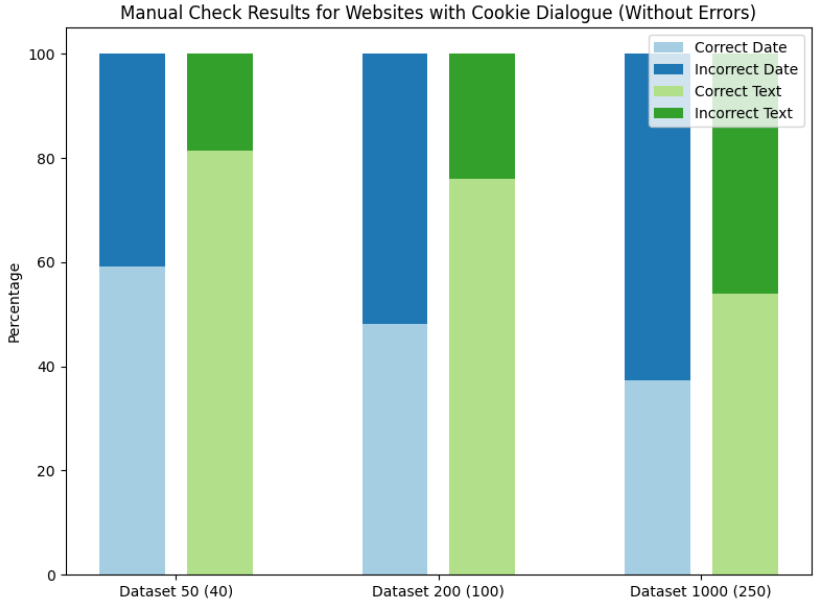


Figure 5.8: These bar charts represent the results of manual checks for three data sets on the websites that have a cookie dialogue without Errors.

These findings are summarized in Table 5.1 for a clearer overview. As the number of websites checked increases, the accuracy of identifying the correct date steadily declines: from nearly 59.5% for the Tranco Top 50 websites to 37.5% for 1,000 websites. A similar pattern is observed with the correct text recognition, dropping from 81.5% to 54%. This decline could be attributed to the complex design of web elements on less popular websites, making it more difficult for the model to identify cookie dialogues accurately.

Table 5.1: Correct Date and Correct Text (Dialogue Present)

Sample Size	**Correct Date (Dialogue Present)**	**Correct Text (Dialogue Present)**
40 (List of 50)	59.26%	81.48%
100 (List of 200)	48.14%	75.92%
250 (List of 1,000)	37.39%	53.91%

Furthermore, we observed discrepancies in how changes in cookie dialogue content were tracked. In 18 instances, changes in the structure of the website led to a misclassification. For example, a single cookie dialogue was split into two distinct web elements that were evaluated

independently by the crawler. This design modification resulted in incomplete records of cookie dialogues, with the crawler incorrectly marking these instances as content changes.

- **General Accuracy of Correct Date:**

$$\text{Accuracy} = \frac{\text{Correct Date (Dialogue Present)} + \text{Correct Date (No Dialogue)}}{\text{Sample size without errors}} \times 100$$

- **General Accuracy of Correct Text:**

$$\text{Accuracy} = \frac{\text{Correct Text (Dialogue Present)} + \text{Correct Text (No Dialogue)}}{\text{Sample size without errors}} \times 100$$

To compute the accuracy of the method, we have taken both **Dialogue Present** and **No Dialogue** results of manual check. This allowed us to evaluate the accuracy of the method performance on three data sets as show on Table 5.2. The accuracy of, on average, 68.18% was shared between all data sets for correctly specifying either the date of fist dialogue occurence or absence of one in case of no dialogue. Undoubtedly, the second case has increased the general accuracy compared to the Table 5.1, but still shows promising results for avoiding false negatives. We also see this effect on the accuracy of cookie content text recognition, which for the list of 50 and 200 websites is on average 85.29% which a deviation of 76.85% for the list of 1,000.

Table 5.2: General Accuracy and Error Rate

Sample Size	**General Accuracy (Correct Date)**	**General Accuracy (Correct Text)**	**Error Rate**
40 (List of 50)	68.57%	85.71%	12.5%
100 (List of 200)	67.44%	84.88%	14%
250 (List of 1,000)	68.55%	76.85%	8.4%

5.6 Discussion

We have tested the proposed methodology through this case study to get a practical idea of its effectiveness and possible limitations. In this section we want to discuss the implications of our findings, acknowledge the limitations, and propose potential improvements to refine the methodology.

The results demonstrate that the methodology is effective in highlighting correlations between key data protection events, like GDPR enforcement, and changes in cookie dialogues. This provides insights into how websites respond to major data protection regulations, which could eventually help measure the broader impact of these regulations on data privacy practices.

However, the study reveals limitations that affect data accuracy. For example, discrepancies in WayBack Machine API snapshots and JavaScript content not loading fully impacted data reliability. Addressing these issues could involve avoiding headless browsing or implementing additional checks for JavaScript presence. Another limitation was due to anti-bot protections

on some websites, which restricted access to essential content. A more sophisticated scraping approach might enhance data integrity and comprehensiveness.

Additionally, we encountered issues with missing records on less popular websites. Filtering out sites with sparse records could improve accuracy, though it may reduce sample size. We also identified challenges in classifying cookie buttons correctly. The implementation failed to classify cookie buttons such as *Reject All* and *Accept All* due to isolated analysis of web elements, suggesting a need for improved element grouping during scraping.

Finally, there is a trade-off between processing time and thoroughness. Increasing search depth could improve element capture but would extend processing time. Balancing these factors is essential to optimize the methodology for future studies. Overall, while the method shows promise, addressing these limitations could enhance its accuracy and reliability, making it more robust for wider applications.

Chapter 6

Conclusion

This thesis aimed to develop a comprehensive methodology for tracking the evolution of cookie dialogues in response to data protection regulations. The goal was to create a scalable, multi-lingual approach capable of analyzing vast datasets over time.

Through the case study of French websites, this research has demonstrated how these dialogues have adapted over time to comply with regulations like the GDPR and CNIL guidelines. By leveraging the WayBack Machine and the XLM-RoBERTa classification model, this study successfully captured patterns and trends in cookie dialogue adoption, offering insights into how data privacy practices evolve in the face of regulatory pressures.

The findings highlight the correlation between enforcement events—such as GDPR implementation and subsequent fines — and the timing of cookie dialogue adoption. This suggests that major regulatory milestones prompt significant shifts in compliance behavior, underscoring the efficacy of data protection laws in shaping the online privacy landscape. However, the study also reveals the variability in adaptability levels across websites, with prominent domains adapting more quickly and thoroughly compared to smaller or less popular sites.

Despite these insights, the study encountered several limitations. Issues with accurately classifying cookie dialogues and variations in website structure presented challenges. These limitations suggest the need for further refinement of the crawling and classification methods, particularly to improve accuracy in identifying and categorizing cookie buttons and dialogue structures. The study also highlights the constraints imposed by the WayBack Machine's archival quality and access restrictions, which could be addressed through future enhancements in web scraping techniques and access strategies.

In conclusion, this thesis's main contribution to longitudinal studies on cookie dialogues includes offering a scalable approach for examining how data protection regulations influence online privacy practices. Future work could enhance this methodology by addressing the identified limitations and exploring more diverse datasets, enabling a broader understanding of privacy compliance trends across different regions and regulatory contexts. By refining these tools, researchers can continue to monitor and analyze the evolving intersection of technology, regulation, and user privacy in the digital age.

Future Work . We believe that future work can enhance cookie dialogue analysis through several key improvements. First, refining the classification model to more accurately identify specific cookie button types, such as "Accept All" or "Reject All," would greatly improve precision. This could involve fine-tuning XLM-RoBERTa or exploring other models specifically trained on diverse cookie dialogues.

Expanding the dataset to include a wider range of websites, including smaller businesses and less popular sites, would provide a broader view of compliance trends. Analyzing websites across multiple EU countries would also help highlight regional variations in GDPR and local DPAs enforcement. Moreover, improvements to web crawling methods, like using multiple archival sources and handling dynamic content more effectively, would enhance data accuracy.

Lastly, future studies could explore the evolution of cookie dialogue language and structure, offering insights into how websites adapt to compliance demands over time. By addressing these areas, future research can deepen our understanding of digital privacy and the effectiveness of data protection laws.

Bibliography

[CKG+20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. 2020.

[Dau23] Carla Dausend. The impact of the gdpr on german online behavior: An analysis of traffic, cookie compliance and online harms. 01 2023.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[DUL+19] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy ... now take some cookies. *Informatik Spektrum*, 2019.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[GGGMO20] Boni García, Micael Gallego, Francisco Gortázar, and Mario Munoz-Organero. A survey of the selenium ecosystem. *Electronics*, 9(7), 2020.

[HIK19] Saad Sajid Hashmi, Muhammad Ikram, and Mohamed Ali Kaafar. A longitudinal analysis of online ad-blocking blacklists. 2019.

[JCNS+22] Hadi Jahanshahi, Mucahit Cevik, José Navas-Sú, Ayşe Başar, and Antonio González-Torres. Wayback machine: A tool to capture the evolutionary behavior of the bug reports and their triage process in open-source software systems. *Journal of Systems and Software*, 2022.

[KJK22] Benjamin Krumnow, Hugo Jonker, and Stefan Karsch. How gullible are web measurement tools?: a case study analysing and strengthening openwpm's reliability. In *Proceedings of the 18th International Conference on Emerging Networking EXperiments and Technologies*, New York, NY, USA, 2022. Association for Computing Machinery.

[KKP15] B T Sampath Kumar, D Vinay Kumar, and K R Prithviraj. Wayback machine: Reincarnation to vanished online citations. *Program: electronic library and information systems*, 2015.

[KP15] B T Sampath Kumar and K R Prithviraj. Bringing life to dead: Role of wayback machine in retrieving vanished urls. *Journal of Information Science*, 2015.

[LOG+19] 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.

[LPVGT+19] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. 2019.

[PSG19] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert?, 2019.

[RSR+23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[SDCW20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[SNGS20] Than Htut Soe, Oda Elise Nordberg, Frode Guribye, and Marija Slavkovik. Circumvention by design - dark patterns in cookie consents for online news outlets. *arXiv*, 2020.

[TLAL22] Kian Tan, Chin-Poo Lee, Kalaiarasi Anbananthen, and Kian Lim. Roberta-lstm: A hybrid model for sentiment analysis with transformers and recurrent neural network. *IEEE Access*, 2022.

[UDF+19] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (un)informed consent: Studying gdpr consent notices in the field. 2019.

[VHASGS22] Marieke Van Hofslot, Almila Akdag Salah, Albert Gatt, and Cristiana Santos. Automatic classification of legal violations in cookie banner texts. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 287–295, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[ZHCG21] Ran Zhuo, Bradley Huffaker, KC Claffy, and Shane Greenstein. The impact of the general data protection regulation on internet interconnection. In *Proceedings of the 2019 ACM SIGCOMM Conference*, 2021.

Appendix A

Legislative Timeline

The purpose of this appendix is to present a constructed timeline highlighting the critical legislative records and authoritative decisions shaping the discussion around cookie consent in France.

Our timeline starts with the ePrivacy Directive (ePD), an early legislative effort to address privacy concerns in the digital environment, particularly focusing on informed consent for cookies. Although our research does not involve investigating for this period, understanding the ePD provides context for subsequent regulations. Moving forward, we examine the General Data Protection Regulation (GDPR), a comprehensive set of laws that changed how personal data can be used and issued specific rules regarding cookie consent more explicitly. This regulation significantly impacted cookie consent mechanisms, necessitating redesigns of consent interfaces to comply with stricter requirements. Lastly, we explore a significant decision by the French Data Protection Authority (CNIL) in September 2020, which refined standards around cookie consent mechanisms in France.

Unimplemented 2019 CNIL Legislation on "Reject All" Option
In 2019, the French Data Protection Authority (CNIL) proposed legislation that mandated a "reject-all" option for cookies, aiming to enhance user consent autonomy and ensure that rejecting cookies was as straightforward as accepting them. This proposal was designed to fully align with the General Data Protection Regulation (GDPR) principles of clear and affirmative consent. However, the legislation never came into effect, possibly due to pushback from industry stakeholders or challenges related to practical implementation [1]. This example underscores the dynamic nature of regulatory efforts in digital privacy and highlights the complexities involved in enforcing such laws.

As we have seen, the journey of digital privacy regulations within France has been marked by continuous adaptations and enhancements aimed at strengthening user consent mechanisms. However, not all proposed changes have been straightforward to implement.

[1] CNIL revised cookie guidelines.

Legislation	Details
ePrivacy Directive (ePD)[2]	• **Date of Directive:** 25 November 2009 • **Legislation:** Directive 2009/136/EC (amending Directive 2002/58/EC) • **Implementation Deadline for EU States:** 25 May 2011 • **Key Requirement:** Introduction of informed consent for cookies, differentiating between required and optional cookies. • **Impact and Insufficiencies:** – Mandated informed consent for cookies, especially those not strictly necessary for service delivery. – Vague definitions of consent led to variable interpretations among EU countries. – France initially adopted a lenient interpretation, where continued browsing was often seen as consent. – This approach was criticized for not offering a genuine choice and potentially overstepping privacy rights. – Resulted in inconsistent implementation across EU member states.
General Data Protection Regulation (GDPR) [3]	• **Date of Regulation:** 25 May 2018 • **Legislation:** Regulation (EU) 2016/679 • **Adaptation Period:** Two-year transition period from adoption on 27 April 2016. • **Key Requirement:** Specification of requirements for active and informed consent. • **Impact and Insufficiencies:** – Introduced a strict definition of consent—freely given, specific, informed, and unambiguous. – Required extensive redesigns of cookie consent mechanisms to comply with new requirements. – Posed challenges for many businesses in France, struggling to balance user experience with legal compliance. – Initial compliance efforts varied widely, reflecting the broad impact and significant adjustments required by GDPR.
CNIL Rulings [4]	• **Date of Regulation:** September 2020 • **Legislation:** Ruling by the French Data Protection Authority (CNIL) • **Adaptation Period:** Implementation required by 31 March 2021 • **Key Requirement:** Mandate of a "reject-all" option for cookies, alongside an "accept-all" option, and prohibition of cookie walls. • **Impact and Insufficiencies:** – Addressed the imbalance in cookie consent mechanisms by mandating a "reject-all" option. – Enhanced user autonomy and privacy by making it as easy to refuse cookies as to accept them. – Posed technical and design challenges for website operators in implementing compliant cookie consent interfaces. – Clarified expectations for consent mechanisms, promoting user-centric consent processes.

Table A.1: Timeline of Major French Legislative Developments Impacting Cookie Dialogues

[2] ePD 2009.
[3] GDPR 679/16.
[4] CNIL regulation.