# Transparency in Privacy Policies

Bianca Bartelt
Fraunhofer Institute for Computer
Graphics Research (IGD), Darmstadt, Germany
Email: bianca.bartelt@igd.fraunhofer.de

Erik Buchmann
Dept. of Computer Science, Leipzig University
Center for Scalable Data Analytics and Artificial
Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
Email: buchmann@informatik.uni-leipzig.de

***Abstract*—Privacy policies are a fundamental concept of the General Data Protection Regulation (GDPR). A privacy policy informs the customers how the organization collects, uses, stores, and shares personal information, and which privacy rights exist. However, a privacy policy can only fulfill this purpose if it is transparent for the customer. In this paper, we analyze 534 privacy policies from the German Top-100 web shops over 7 years, starting in 2016. We want to find out whether changes in data protection regulations and related events had an impact on transparency in privacy policies. Furthermore, we want to compare our results with international findings. We define transparency as readability and discoverability of mandatory information. We observed that the GDPR has increased the length of German privacy policies, but also the discoverability of mandatory information. However, the GDPR has not made the policy texts easier to read. This is in line with international studies, that used a different approach to analyze transparency.**

***Index Terms*—Privacy Policy, Transparency, GDPR**

## I. Introduction

Privacy policies are essential for any digital service. Such a policy demonstrates the service's commitment to the responsible handling of customer data by explaining how personal data is collected, used, stored and/or shared with another company or organization. It empowers customers to control their data, by outlining fundamental customer rights, such as the right to access, correct, or delete information. Finally, privacy policies allow the customers to assess their risks associated with data breaches, misuse, or unauthorized access.

The content of a privacy policy depends on the organization, its activities, and legal requirements of the country it operates in. Important information for the customers can be represented in various ways and different places. For this reason, the General Data Protection Regulation (GDPR) [1] defines transparency as a requirement for privacy policies.

We want to quantify how transparency has changed in the privacy policies of web shops over time. Transparency has many different aspects and definitions [2]. We define it as (a) the *readability* of the text body of the policy and (b) the *discoverability* of information that are mandatory for privacy policies according the GDPR. Other aspects of transparency, e.g., correctness, completeness or the use of technical terms, would require internal knowledge of the organization that published the policy, or do not make sense for privacy policies.

Anecdotal evidence indicates that privacy policies have become longer and more structured in the last years, but it remains unclear whether this has led to more or less transparent policies. In order to explore this, we have collected 534 privacy policies from the German Top-100 web shops from 2016 to 2022, and we analyzed them with text statistics and Natural Language Processing (NLP).

The work closest to us is [3], which investigates the impact of the GDPR on EU privacy policies from 2016 and 2019. With 470 participants from Amazon mechanical turk, the study manually assessed the visual improvement of privacy policies after the GDPR became active. It shows that in general, the visual appearance of EU privacy policies has indeed improved from a subjective user perspective. However, the EU consists of many regions with distinct cultural and juridical attitudes. Thus, we strive to find out whether this also applies for German privacy policies, and whether such an analysis can be automated. We make the following contributions:

- We describe an approach to fetch and clean privacy policies from public sources, such as the Internet Archive.
- We measure transparency as *readability* and *discoverability*, using readability metrics and NLP.
- We compare our findings with previous work on international privacy policies.

We learned that the GDPR has increased the median length of the policies by 325%, but did not reduce the high demands on the reading skills required. However, the GDPR had a positive impact on the discoverability of important information. In comparison with international studies, we found that the GDPR indeed not only standardized the content of the privacy policies, but also the readability and discoverability.

**Paper structure:** Section II reviews related work. Section III outlines data selection and cleansing. The Sections IV and V analyze readability and discoverability. Section VI summarizes our findings, and Section VII concludes.

## II. Related Work

This section introduces transparency, readability and NLP.

### A. Transparency and the GDPR

Art. 5 GDPR [1] and the corresponding Recital 39 state that lawful and fair processing of personal data requires transparent information to the persons concerned, and makes transparency a fundamental principle. Recital 58 defines transparency as "*any information* $(\cdots)$ *be concise, easily accessible and easy to understand, and that clear and plain language* $(\cdots)$ *be used.*" Thus, it is important, that the information is presented in a clear language. It must also be easy to find, without browsing the small print of a lengthy privacy policy.

Art. 12ff. GDPR specifies which information must be made transparent. This is (i) how personal data is collected and handled, and (ii) the privacy rights of the data subject. (i) includes which categories of data are collected, the purpose of processing of the collected data and the time after which the data is deleted. If the data is transferred to third parties, the recipients must be made transparent. In addition, contact information of the organization, its representatives and (if present) its privacy officer must be made public. (ii) includes the rights to access (Art. 15), to rectification (Art. 16), to erasure (Art. 17), to restriction of processing (Art. 18), to be notified (Art. 19), to data portability (Art. 21), to object (Art. 22) and to involve an authority (Art. 77). If the data processing depends on a consent, it must be clear how to withdraw the consent (Art. 7).

### B. Text Features and Readability Metrics

Understandability is an important aspect of transparency. Readability metrics quantify understandability by calculating an index value from statistical text features. Such features are different for each language. Thus, the parameters are gauged for each language, based on reference texts. The Flesch-Reading-Ease (*FRE*) maps the average length of a sentence $ASL$ and the average number of syllables per word $ASW$ to a scale: $FRE = 180 - ASL - (ASW \cdot 58.5)$. Numbers below 30 indicate texts that are very difficult to read [4]. The 4th. "Wiener Sachtextformel" (*WSF*) has been specifically designed for German texts. Besides ASL, it also considers the percentage of three- and polysyllabic words (MS) [5]: $WSF = 0.2744 \cdot MS + 0.2656 \cdot ASL - 1.693$. The resulting value represents a reading competence between the 4th and 15th grade of a (hypothetical German) school, i.e., low numbers indicate better readability.

TABLE I
READABILITY METRICS.

| Readability | FRE | WSF |
|---|---|---|
| very hard | 0-30 | 13-15 |
| hard | 30-50 | 12 |
| rather hard | 50-60 | 11 |
| medium | 60-70 | 9-10 |
| rather simple | 70-80 | 7-8 |
| simple | 80-90 | 6 |
| very simple | 90-100 | 4-5 |

Table I maps the index values to reading competences. Other metrics produce comparable results, e.g., the Lasbarhetsindex [6] or the Gunning Fog Index [4].

### C. Natural Language Processing

A broad range of NLP technologies has been proposed to automatically process natural language [7] [8]. Typically, the first processing step of NLP is *Tokenization* [9], which separates the text body into entities, such as words, punctuation, dates or symbols. Those tokens can be further processed, e.g., with *Part-of-Speech* (PoS) tagging [10]. PoS tagging assigns labels to tokens that tell if a token is a noun, verb, adverb, conjunction word, etc. PoS taggers use models that are gauged or trained for the sentence structure of a language. *Lemmatization* [11] derives the base forms of words, e.g., translates a token "Transferral" to "transfer". *Stemming* [12] goes beyond that, by removing and replacing suffixes to obtain the root form of a word. The root of "Transferral" would be "transferr". Both approaches can be used to normalize a text.

The *term frequency–inverse document frequency* (TF-IDF) quantifies how well a word distinguishes a certain text from a set of other texts [13]. For example, the token "privacy" appears frequently in all privacy policies, and does not allow to tell them apart. But it might be suitable to distinguish a privacy policy from security specifications. TF-IDF computes the relative frequency of a term within a document, multiplied with the logarithmically scaled inverse fraction of the documents that contain the term.

## III. DATA SELECTION AND CLEANSING

Our research objective is to assess the development of transparency in privacy policies. We are interested to see whether changes in the legislation or events with a large impact on the customer's attitudes towards privacy have led to more or less transparent policies. Such events include the EU-wide activation of the GDPR in 2018 [1], the cancellation of the EU-US Privacy Shield [14] in 2020, and in Germany the data leakages connected with Facebook and Covid 19 tests in 2021 [15]. Furthermore, we want to find out to which extent it is possible to do this automatically, to evaluate a large number of policies. Our research approach consists of five steps: *Data Collection, Data Cleansing, Readability Analysis, Discoverability Analysis* and *Transparency Evaluation.*

### A. Data Collection

Observations show that the perception of privacy risks and the everyday implementation of privacy regulations depend on the society. We focus on privacy policies from German web shops, because we are familiar with the social events and trends necessary to interpret our analysis results. We are interested in policies from 2016-2022. This allows us to compare our findings with an international study [3], which analyzed policies from 2016 and 2019. In this study, participants from Amazon mechanical turk manually assessed privacy policies collected in the EU, i.e., across different societies.

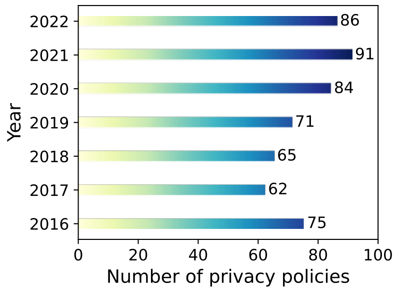

Figure 1. Privacy policies per year.

To fetch the privacy policies from different years, we implemented a Python script that downloads policies not only from the Internet, but also from the Internet Archive [16]. An input parameter of this script is a list of links to privacy policies. We obtained this list from a Top-100 ranking of German business-to-consumer web shops with the highest revenues [17]. Because some web shops have modified the URLs to their privacy policies over the years or generate the policy depending on user interactions on the fly, our script failed to download some policies. Figure 1 shows the number of downloaded policies per year. We obtained a complete set of policies over the entire study period for 32 web shops. For 28 further web shops, the policy from one year of the study period could not be downloaded. In total, we obtained a body of 534 policies.

### *B. Data Cleansing*

The privacy policies were downloaded in a HTML format. They must be cleansed and filtered for further analyses. We used the Python library "BeautifulSoup" [18] to parse the HTML code, and we removed header, footer, navigation bars, menu entries and any HTML tags that are not necessary to analyze the structure and contents of a privacy policy. This reduced the size of the data set by 70%. Figure 2 provides a running example of such a cleansed privacy policy.

```
<body>
    <div id="content">
        <h1>1. Data We Are Gathering About You</h1>
        When you sign up for a service, we collect your
            contact information. We collect and use account
            data to process payments.
    </div>
</body>
```

Figure 2. Cleansed privacy policy.

Note that we we have translated this example to English for better understanding – our policies were written in German. Further processing steps need to be different for the assessment of readability and discoverability.

## IV. Readability of Privacy Policies

In this section, we want to quantify the readability of the policies with text statistics and readability metrics. We therefore need to pre-process our data set. We filtered out any text that is not a full sentence, e.g., headlines and enumerations, by using BeautifulSoup and regular expressions.

```
When you sign up for a service, we collect your contact infor–
mation. We collect and use account data to process payments.
```

Figure 3. Pre-processed policy for readability analyses.

Figure 3 illustrates the result with our running example. This reduced the size of the downloaded data set to 42%. Some policies were not parsable, e.g., due to texts generated by JavaScript. This resulted in texts too short for a meaningful statistical analysis. We removed any policy from our data set that had less than 500 characters left. A threshold of 500 characters corresponds to 6-8 German sentences on average. At the end, we obtained 439 cleansed, pre-processed and filtered policies for our analyses.

We start with text statistics. In general, a policy that contains the same mandatory information with less or shorter words and less or shorter sentences is more readable and therefore more transparent, than a long text with long words. Figures 4 and 5 show the number of words and the number of sentences of our privacy policies (without headlines and enumerations) for each year between 2016 and 2022. The boxes show the 25% and 75% quartiles and the median values. The whiskers end at last value equal or smaller than 1.5 interquartile range. Dots represent outliers.

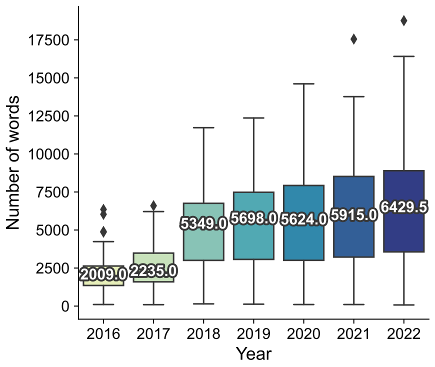

Figure 4. Words.

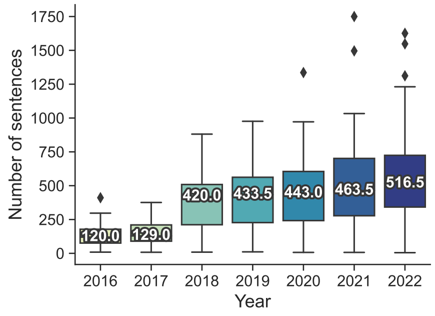

Figure 5. Sentences.

The figures show a significant increase both in the number of words and the number of sentences in 2018, the year of the activation of the GDPR. The median number of words more than doubled, the median number of sentences more than tripled. We also see that the spread between the first and third quartile increased, presumably due to different interpretations of the level of detail needed to publish mandatory information. We also see a smaller increase in the values in 2021. In this year, the cancelled EU-US Privacy Shield [14] had to be replaced. Furthermore, 2021 was notorious in Germany for its data privacy scandals [15].

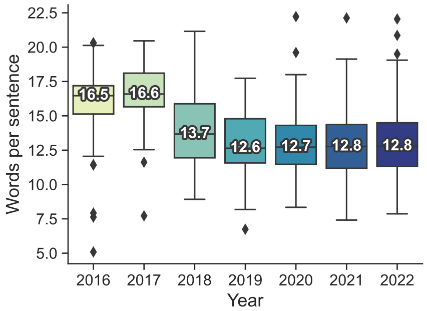

Figure 6. Words per sentence.

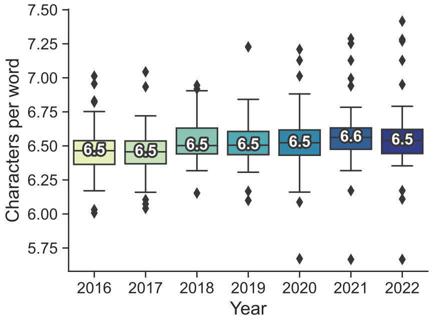

Figure 7. Characters per word.

Typically, German public media use 12-15 words per sentence, and 6-8 characters per word on average [19]. Figures 6 and 7 show the number of words per sentence and the number of characters per word for our policies. To our surprise, the median of the number of characters per word did not change over years, and remained in the typical range for public media. Furthermore, we observed that the activation of the GDPR resulted in shorter sentences, which are more readable than long sentences. Observe that the median number of sentences rises each year to a similar extent as the median number of words per sentence falls, i.e., even if the policies tend to grow, the length of the sentences decrease.

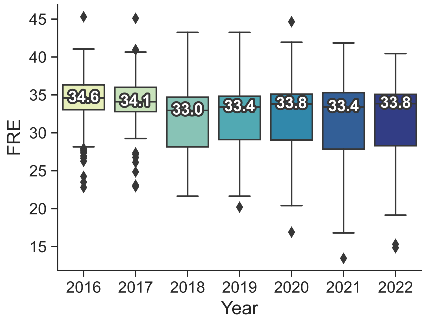

Figure 8. Flesch Reading Ease.

Finally, we compute the readability metrics. Figure 8 shows the FRE for our policies. We have observed similar results with the “Wiener Sachtextformel”. Both metrics were calculated with the “Textstat” library [20]. Surprisingly, the median of the required reading skills hardly changes over the years. According to Table I, the FRE indicates that the median policy is “hard” to read, close to “very hard”. A FRE below 30 corresponds to the reading competence of an academic.

In addition, the range between the 25% and the 75% quartiles increased with the activation of the GDPR. In particular, the lower quartile expanded significantly. Since lower numbers indicate less readable texts, this means that with the activation of the GDPR, a significant share of the policies have been rewritten in a less readable, less transparent manner. This situation has even worsened in 2021.

## V. Discoverability

In this section, we analyze the discoverability of mandatory information in privacy policies. Recall that we do not assess correctness or completeness. Table II lists the six classes of information we are focusing on (cf. Section II):

TABLE II
Classes of information.

| Class | Description |
|---|---|
| 1 | Storage period |
| 2 | Categories of data |
| 3 | Purpose of processing |
| 4 | Recipients of data |
| 5 | Contact information |
| 6 | Rights of the data subject |

This information must not be hidden within the text body, but highlighted to some extent. Thus, we pre-process our policies by filtering the HTML tags for headlines (*h1, h2, h3, h4, h5, h6*) and bold texts (*b, strong*) with BeautifulSoup, as shown in Figure 9. Note that there might be options to highlight text in HTML, which we cannot recognize with this procedure, e.g., by using JavaScript or style sheets to create or re-purpose tags that are non-highlighting by default.

```
1. Data We Are Gathering About You
```

Figure 9. Pre-processed policy for discoverability analyses.

We did not find pre-trained NLP classifiers that produce accurate results on German privacy policies. Thus, we had to build our own classification toolchain. Therefore, we tokenized the highlighted lines of text first, and we used Part-of-Speech tagging [10] from the Textstat library [20] to remove any characters except nouns and verbs.

```
data; gather
```

Figure 10. After tokenization, filtering and lemmatization.

We reduced the remaining words to their basic forms (Lemmatization), and removed duplicats if present. Figure 10 illustrate the result of this pre-processing. Finally, we removed any words with no meaning for our classification, e.g., “privacy

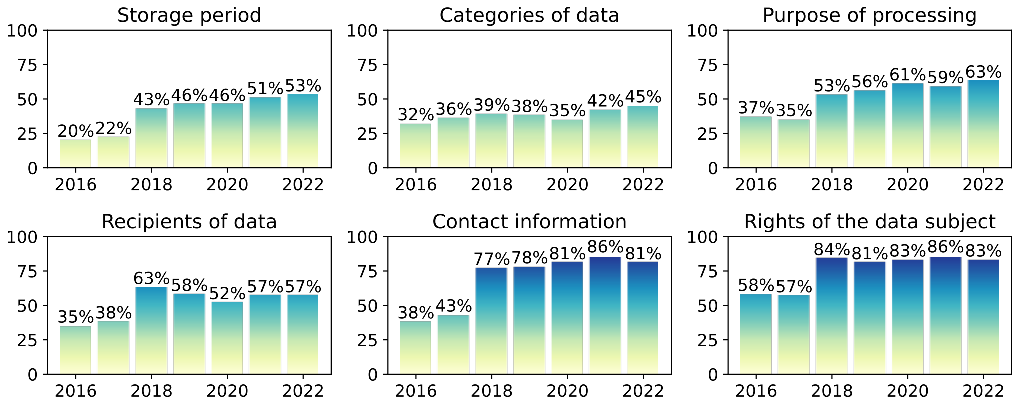

Figure 11. Discoverability of mandatory information.

policy" or "cookie". In total, we obtained 327,782 highlighted lines of text with an average length of 27 characters.

To classify our tokenized, filtered and lemmatized lines, we manually generated a reference data set of the 250 most frequent combinations of tokens. We labeled them manually with the classes shown in Table II. For example, {data; gather} would be labeled with "Categories of data", because such a headline in a privacy policy announces an explanation on the data collected. We used the Python library "scikit-learn" [21] for the classification. In particular, we computed the cosine similarities between the term TF-IDF vectors of our labeled reference data set and our highlighted lines of texts. We assigned a highlighted text with a class label, if the cosine similarity set was better than 0.75. With this threshold, we obtained a classification accuracy of 90%.

Figure 11 shows the percentage of classes identified in the highlighted lines of our 439 privacy policies. A 86% in Figure "Contact information" for 2021 means, that in 86% of all inspected privacy policies fom 2021 a highlighted line of text exists that addresses options to contact the data collector. The numbers represent lower bounds, because we do not have 100% accuracy and cannot find texts highlighted with unusual tags. Nevertheless, it shows tendencies:

Because the GDPR requires to make certain information visible, there is a significant increase in 2018 over all classes. There exist classes of information that are easy to find in almost all privacy policies. In particular, this is contact data and the enumeration of the rights of the data subject. These rights can be copied from the GDPR and are the same for every organization. Observe that policies from before 2018 had to grant corresponding rights from the national predecessor of the GDPR ("Bundesdatenschutzgesetz" [22]).

There are classes of mandatory information that potentially conflict with the business interests of a company, e.g., for how long personal data is stored and which categories of data are collected. Figure 11 shows that such information indeed is less likely to be discoverable in a German privacy policy.

Figure 12 provides a heatmap of discoverable information. The fields count the privacy policies that show a certain number of mandatory information easily discoverable in a certain year. For example, consider the "18" in the field for 3 classes of discoverable information in 2020. This means that from our body of downloaded privacy policies a total of 18 announced 3 different classes of mandatory information in a highlighted part of the policy text.

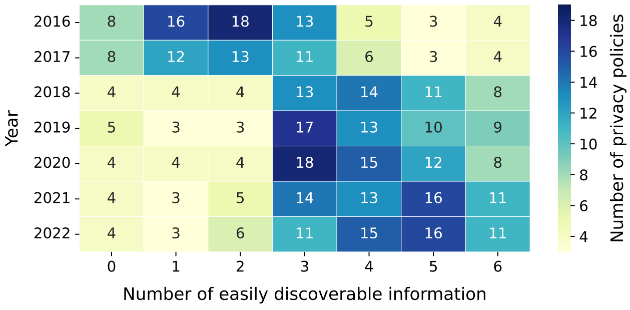

Figure 12. Discoverable information.

The figure shows an increase of +18% to +34% in discoverable information due to the activation of the GDPR, and an ongoing tendency to highlight more information.

## VI. Transparency in Privacy Policies

In this section, we discuss our findings on transparency on German privacy policies, and we compare them with international studies.

We have defined transparency for privacy policies from the external perspective of the intended auditory. From this perspective, transparency can be understood as readability and discoverability of information on data handling practices and on rights of the data subject. The GDPR makes it mandatory for each organization to lay open such information, if personal data is handled.

Our readability analysis has shown, that German privacy policies require a reading competence on high school level. The effect of the GDPR was that the required reading level for a part of policies went to the level "academic", while the median level did not change, and the policies generally became much longer. We also see an increase in the length of the privacy policies in 2021. In this year, companies had to circumvent the cancelled EU-US Privacy Shield [14] and many data privacy scandals [15] gained public attention. Since this had no impact on the reading levels, we assume that additional statements were added to the policies that specifically address these privacy issues.

Our discoverability analysis is more positive. It shows that the GDPR has led to a large increase of the discoverability of important privacy rights. Even the discoverability of information that might be in conflict with business interests has increased. The effect of other privacy-related events on discoverability seems to be negigible. We conclude that the GDPR makes a very positive contribution to the transparency of privacy policies, and that organizations indeed react on privacy issues that are of concern for its customers.

Past studies analyzed privacy policies from English-speaking countries for their length [23], readability [24] [25] or content and visual appearance [3]. The studies have a broad focus, i.e., do not specifically analyze transparency and consider a mixed set of privacy policies that includes different cultural and juridical attitudes. The studies show that the length of English policies also increased significantly due to the activation of the GDPR. Similarly, the studies observed that policies in general are difficult to read. We observed a FRE median value of 33.4. A readability analysis [23] reports an average FRE of 32.8 in the EU for 2018, and points out that the average FRE for privacy policies worldwide in the same year is 39.8 (still level “hard”, but slightly less demanding).

To the best of our knowledge, an international study on discoverability does not exist. However, participants of a study [3] found that the visual appearance of privacy policies has been improved with the GDPR. This might indicate that important information are easier to find, so that the policies could be more transparent. We conclude that the GDPR not only successfully standardized the content of privacy policies through Europe, it also ensured that privacy policies became more transparent, regardless of the languages used.

## VII. Conclusion

The General Data Protection Regulation requires that privacy policies declare the handling of personal data and the rights of the data subject in a transparent way. We have defined transparency as (a) the *readability* of the policy texts and (b) the *discoverability* of mandatory information in the policy. To this end, we have analyzed 434 privacy policies from the German Top-100 web shops from 2016 to 2022 with text statistics, readability metrics and natural-language-processing.

We observed that the GDPR has increased the median number of sentences by 325%. The GDPR increased the discoverability of mandatory information, but it did not reduce the high demands on the reading skills required. By comparing our findings with studies on privacy policies in the EU and globally, we found that the GDPR has not only standardized the content of privacy policies, but also aligned their readability and the information that needs to be highlighted.

## References

[1] European Union, “REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” Official Journal of the European Union, L119/1, 2016.

[2] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamo-Larrieux, “Robots and transparency: The multiple dimensions of transparency in the context of robot technologies,” *IEEE Robotics & Automation Magazine*, vol. 26, no. 2, pp. 71–78, 2019.

[3] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, “The privacy policy landscape after the GDPR,” *arXiv preprint arXiv:1809.08396v3*, 2019.

[4] E. Robinson and D. McMenemy, “To be understood as to understand,” *Journal of Librarianship and Information Science*, vol. 52, no. 3, pp. 713–725, 2020.

[5] R. Bamberger and E. Vanecek, *Lesen-Verstehen-Lernen-Schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache*. Jugend und Volk, 1984.

[6] J. Anderson, “Analysing the Readability of English and Non-English Texts in the Classroom with Lix.” in *Proceedings of the Annual Meeting of the Australian Reading Association*, 1981.

[7] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: an introduction,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[8] P. Johri *et al.*, “Natural language processing: History, evolution, application, and future work,” in *Proceedings of the International Conference on Computing Informatics and Networks*. Springer, 2020, pp. 365–375.

[9] J. J. Webster and C. Kit, “Tokenization as the initial phase in NLP,” in *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.

[10] A. Chiche and B. Yitagesu, “Part of speech tagging: a systematic review of deep learning and machine learning approaches,” *Journal of Big Data*, vol. 9, no. 1, pp. 1–25, 2022.

[11] R. Sonbol, G. Rebdawi, and N. Ghneim, “The use of NLP-based text representation techniques to support requirement engineering tasks: A systematic mapping review,” *IEEE Access*, 2022.

[12] V. Balakrishnan and E. Lloyd-Yemoh, “Stemming and lemmatization: A comparison of retrieval performances,” *Lecture Notes on Software Engineering*, vol. 2, pp. 262–267, 2014.

[13] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2011.

[14] European Union, “Commission Implementing Decision (EU) 2016/1250 on the adequacy of the protection provided by the EU-U.S. Privacy Shield,” Official Journal of the European Union, L207, 2016.

[15] D. Kolaric, “Top 5 der größten deutschen Datenskandale 2021,” https://www.all-about-security.de/top-5-der-groessten-deutschen-datenskandale-2021/, 2022, [retrieved: Dec., 2023].

[16] Internet Archive, “The Wayback Machine,” http://archive.org/, 2023, [retrieved: Dec., 2023].

[17] EHI Retail Institute, “Top-100 Onlineshops in Deutschland,” https://www.ehi.org/news/top-100-onlineshops-in-deutschland, 2022, [retrieved: Dec., 2023].

[18] L. Richardson, “BeautifulSoup,” https://www.crummy.com, 2023, [retrieved: Dec., 2023].

[19] WORTLIGA Tools GmbH, “Glossar,” https://wortliga.de/glossar, 2023, [retrieved: Dec., 2023].

[20] C. A. Shivam Bansal, “Textstat,” https://pypi.org/project/textstat/, 2023, [retrieved: Dec., 2023].

[21] scikit-learn, “Machine Learning in Python,” https://scikit-learn.org, 2023, [retrieved: Dec., 2023].

[22] Bundesrepublik Deutschland, “Bundesdatenschutzgesetz, in der Fassung vom 14.01.2003 (BGBl. I S. 66), zuletzt geändert durch Gesetz vom 30.10.2017 (BGBl. I S. 3618),” BGBl. I S. 3618, 2017.

[23] I. Wagner, “Privacy policies across the ages: Content and readability of privacy policies 1996–2021,” *arXiv preprint arXiv:2201.08739*, 2022.

[24] B. Fabian, T. Ermakova, and T. Lentz, “Large-scale readability analysis of privacy policies,” in *Proceedings of the International Conference on Web Intelligence*, 2017.

[25] R. Amos *et al.*, “Privacy policies over time: Curation and analysis of a million-document dataset,” in *Proceedings of the Web Conference 2021*, 2021, pp. 2165–2176.