Do they really value your privacy?

An exploratory analysis of what can be learned about companies from their privacy statements

T.P. Rovers

Technische Universiteit Delft

TU Delft Delft University of Technology

Challenge the future

Delft University of Technology

Do they really value your privacy?

An exploratory analysis of what can be learned about companies from their privacy statements

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of

Master of Science
in **Engineering & Policy Analysis**

To be defended in public on October 18, 2019

Author:
Timothy Rovers
Student number: 4211731

Supervisor:
dr. H. Asghari

Graduation Committee:
dr. M.E. Warnier
mr. drs. W. Ernst

October 6, 2019

TU Delft

Deloitte.

Acknowledgements

The document you're currently reading is a product spanning many meetings, iterations, countless hours of Python, some great successes and a healthy dose of frustration and failures. But now it's finished, I'm very happy with the result, and very happy that you'll take the time to read it (or at least part of it). It was approximately a year ago that I e-mailed Hadi, wondering if he was still looking for thesis students. Back then I had absolutely no idea how I would ever complete a master thesis, let alone where to start. That's why I'd like to thank the people who guided me through this whole process, my graduation committee, whom I'm still convinced I couldn't have chosen better.

First of all thank you Hadi. For your great insights, motivation to keep looking for new findings, and patience when I didn't fully understand all your ideas. You were a great supervisor, and your critical questions and creativity have definitely had a large impact on making the thesis what it is. During one of our meetings you said the great sentence *"The great thing about what you've done is that you've done it, so now we can tell you why you're wrong"*, which is not only quite funny, but also some amazing advice for everyone writing a thesis to just start doing things instead of endlessly planning everything.

Second, many thanks to you Martijn, you were a great second supervisor. Whenever I felt I had a problem and came to you, you already understood the problem better than I did and had solved it. It's amazing how you're able to give such structural and useful advice while guiding so many students, and you've definitely also had a great influence on this research.

Also many thanks to you Wouter. It was great to have you as a supervisor at Deloitte, your knowledge about privacy and your law-oriented advice motivated me to keep my eyes open for the problem I was attempting to tackle. You also had some great ideas for me on how to structure my work, were always open for questions, and kept me motivated during the sometimes long days of coding and writing.

Lastly I'd like to express gratitude to my friends and family who supported me in this process and made this possible. A special thanks to my parents, who've always supported me and pushed me to perform in school and through university.

Enjoy reading!

Executive Summary

Background
A growing global challenge which is faced today is that of data privacy. An ever-growing amount of companies is finding ways to collect and use consumer data in order to produce additional revenue, or even as their main source of revenue. The prime example of this is Google, which has made collecting different types of consumer data an essential part of their business model. This data is subsequently used for marketing purposes, often without the knowledge of the consumer. The asymmetry of knowledge between consumer and company is a problem, as consumers are sharing personal information without knowledge of what the intended or even unintended consequences can be, or what the real value of their data is.

In an attempt to give more power to consumers, large efforts are being made in the form of new privacy regulations. It is currently nearly 1.5 years since the General Data Protection Regulation (GDPR) of the European Union came into effect, which has forced all companies to review the way data, and mainly personal data, is collected, used and shared. The GDPR has had positive effects on the transparency of companies and lead to a general improvement of the business-side knowledge of privacy, but the issue mentioned in the first paragraph has not been solved yet. Thus, companies like Google and Facebook still benefit from providing free services to attract as many consumers as possible, which enlarges the amount of data they have. Because of this, research is being done to better understand this problem. The new theory of *Surveillance Capitalism* emphasizes the economic value of data, and shows that as long as data is a commodifiable good it will be collected in large quantities. When no new regulation is pursued, companies will maximize the amount of data they extract. Companies currently doing this are defined as surveillance capitalists.

Research
To gain some further insights into this matter, this thesis attempts to find signals which can indicate the degree to which a company is committed to privacy. These signals are sought for in privacy statements. As described earlier, the GDPR has improved transparency of data usage by obligating companies to share certain information of their data practices in their privacy statements. These statements can therefore potentially be a valuable source of information in understanding how companies handle privacy, and possibly identify companies which are using unlawful, unethical or unfair data practices.

Based on the problem described above, the following problem statement is defined: *"Can a company's commitment to privacy be derived from their website's privacy statement using Natural Language Processing? If so, can other underlying attributes of privacy commitment be derived which could indicate surveillance capitalism? What implications does this have for privacy regulators and governments?"* Because of the scope of the problem and the novelty of the approach, this research is tackled in an exploratory manner. The goal of the research is not to test and validate a proven method, but to test the limits of what can be learned, and find out if certain methods hold promise for future work. This is done in three general steps. A new database of privacy statements is created originating from the most-viewed websites in English. Next, Natural Language processing is used to extract a set of variables from the statements, enumerating aspects of each privacy statement. Lastly, the variables are analyzed in a number of ways, of which the findings partly function as cross-validation by confirming findings with existing literature, and partly as an assessment of

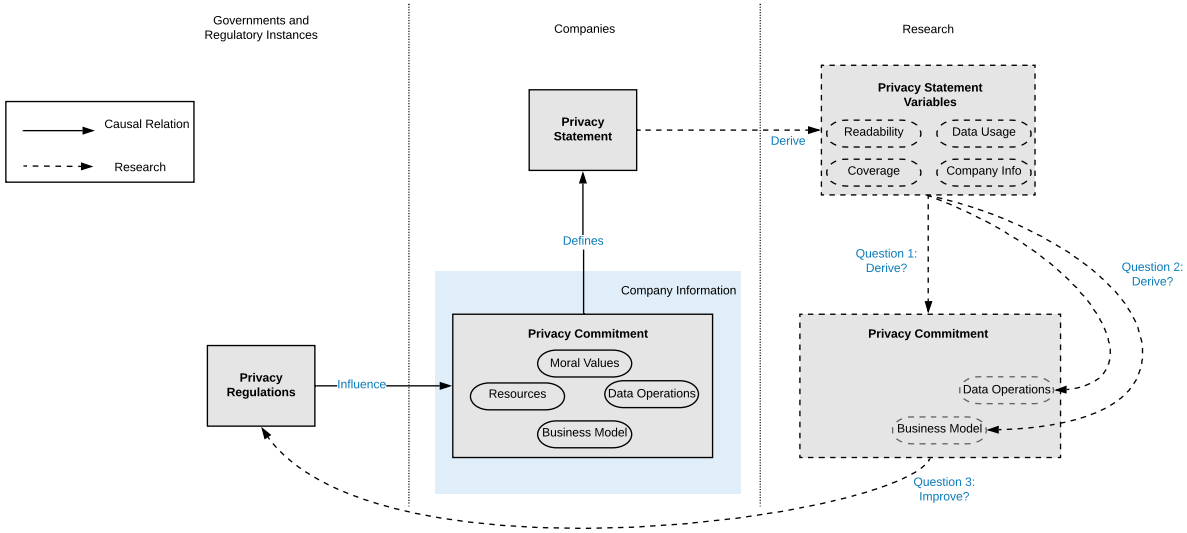


Figure 1: Conceptual Framework

which methods have promise for future work. This will ultimately lead to answering the three questions defined in the problem statement. This process is schematically shown in figure 1.

Method

First, a large database of statements is collected. This is done using Amazon Mechanical Turk, creating a task which people can complete online for a small fee. In this case, the task consisted of visiting a URL (pre-defined as one of the most-visited websites in English), finding and copying the link to the website's privacy statement, entering the date the statement was last updated, and copying the full text of the privacy statement. This was done for 2000 websites and ultimately, after the process was completed and the data was cleaned, resulted in a database of 1508 privacy statements. A selection of these was used to create the natural language processing algorithms which extracted an eventual total of 72 variables. Due to the scope of the project, these variables were largely based on variables defined in existing literature, along with a number of self-created additions. Each variable was tested to have an accuracy of at least 80%, otherwise the variable was not used or revised.

Analyses

Once the dataset was complete, multiple analyses were performed. First, an analysis was performed to test for distinctions between statements from different regions and statements from different categories. From this analysis, a number of significant differences could be found between the groups. A number of these could be cross-validated with other research, indicating that the methods succeeded in extracting useful information. For instance, nearly all of the European statements were updated shortly before GDPR. Also, news websites had a significantly higher presence of third party advertisers, which is in line with expectations (due to their business model) and existing literature. A number of additional findings were done, such as that companies belonging to the website category *Health & Fitness* have the highest presence of the third party Google.

The next analysis performed was an assessment if surveillance capitalists could directly be distinguished from similar companies based on the variables extracted. A comparison was done between privacy statements of the big-5 tech companies, which include Facebook and Google. Interestingly, results showed that the statements of surveillance capitalists were easier to read, less vague, and did not mention sharing many

types of data. This was contrary to initial expectations.

Next, a metric was created based on the variables which could indicate the quality of a privacy statement. The variables were chosen based on existing literature. These metrics were informativeness, complexity and length. Based on these 3 variables, two groups of best- and worst cases were identified, which upon manual inspection gave a good indication of the quality of a statement. This also made it possible to assess which variables were signals between these two groups, showing that discussing choices of the user was the main signal for a good privacy statement.

Following this, a similarity analysis was performed, which made it possible to indicate the percentage of identical sentences between statements. This found that 12% of the sentences occurred more than once, and approximately 100 sentences were copied more than ten times. It also showed that approximately 50% of the statements shared at least one sentence with other privacy statements, with approximately 1% sharing as much as 60% of their contents with a totally different, unrelated company. This analysis therefore indicated that many companies simply copy text from other statements, which in turn indicates a low commitment to privacy. This information and method can be further expanded to create a copying metric, which in turn can be used to further indicate commitment to privacy.

Conclusions and implications

The first achievement of this research was the creation of a database of 1508 privacy policies, which can be used for future research. Next to that, 72 different variables related to privacy where extracted from the privacy statements, which can provide insights into the way a company handles data. This was done using relatively simple natural language processing algorithms while achieving a high accuracy. From this, methodological conclusions can be made on how to extract certain types of privacy related information in a simple manner from privacy statements.

Based on the findings above, it was concluded that it is possible to extract commitment to privacy using the methods in this research. The analyses succeeded in creating a metric which can indicate the quality of a statement and an indication of the amount of sentences which are potentially copied from other statements. These two metrics form indications of commitment to privacy, as they are signals of the effort a company has put into creating their privacy statement. Additionally, more complex variables could form signals of commitment to privacy, which provide opportunities for future work. However, based on the methods used in this research no clear indications of surveillance capitalists could be identified. That being said, the methods applied in this research are elementary, and further research is encouraged on this topic.

The first recommendation that can be made is directed at privacy regulators which are interested in more efficient methods of checking compliance of companies with privacy regulations. For these regulators, the recommendation is made to support the creation of a metric of commitment to privacy based on privacy statements, as using the methods in this research have shown large promise to succeed in doing so. Further research is encouraged on the topic of extracting additional information from privacy statements, as further analysis has potential of uncovering additional signals for identifying commitment to privacy and surveillance capitalism in privacy statements, which could further improve the usability of these metrics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Big data; on one hand it is a term associated with endless technological possibilities, increased knowledge for business and consumer, and unmissable for companies aiming to outsmart competition; on the other hand it is an ambiguous term encompassing an ever-growing amount of data bringing risks of large-scale privacy issues (Ward & Barker, 2013; Acquisti, Taylor, & Wagman, 2015). With the advancement of technology, the size and therefore importance of big-data within our society has grown. A growing number of machines is able to record and collect, process and upload specific data; be it our laptops which track what websites we visit, our smart-watches which track our location or our smart-thermostats which control the lights, security cameras and temperature in our houses (Acquisti et al., 2015; Hernandez, Arias, Buentello, & Jin, 2014). Each device is capable of uploading this collected data to enormous central databases, which combined with more powerful computers running more powerful algorithms provides the possibility of analyzing huge amounts of data and gaining ever-growing amounts of knowledge based on this data, which can be used by these companies to further improve their businesses in a multitude of ways (Tene & Polonetsky, 2012).

The increasing number of possibilities also bring an increasing number of privacy-related risks. Given the technology present, tech-companies are now theoretically able to constantly watch, listen and follow any consumer who uses their devices. In order to prevent this from actually happening, governments and inter-governmental organizations as the European Union (henceforth EU) have been enforcing (data-)privacy laws for a multitude of years. The EU officially acknowledged the issue of privacy since defining it as a universal human right in 1950 (Council of Europe, 1950). The most recent regulation of this type enforced by the EU has been the General Data Protection Regulation (EU Regulation 2016/679, henceforth GDPR), which requires data controllers and data processors (see section 2.1.1 to be more transparent and informative about their data collection and processing practices. Most companies which collect personal data have therefore updated their privacy policies in order to comply with these rules, as non-compliance can result in penalties up to 20 million euros or 4% of the total worldwide revenue (Linden, Harkous, & Fawaz, 2018; European Union, 2016). The GDPR has therefore forced many companies to carefully inspect their data-practices in order to avoid the fines mentioned.

Although providing significant improvements, the GDPR is not the ultimate solution to the growing issue of data-privacy. Shoshanna Zuboff (2015) describes this by defining the economic imperative which is driving many companies to collect an increasing amount of data, which she calls *Surveillance capitalism*. Surveillance capitalism is a form of capitalism where the a company's financial imperative is to maximize the amount of data it receives from it's users, in order to be able to process this data into prediction products. These prediction products can then be sold on behavioral future markets to companies which can use this data to gain insights on users behavior, therefore being able to predict what the users will do. The issue here is that the personal data is claimed, according to Zuboff: surveillance capitalists provide a service in order to acquire as many users as possible. These users provide the data which are used for a surveillance capitalist's actual

business model, namely the processing of this data and the sales of this processed data to third parties.

The GDPR has aimed to solve this problem of the 'claiming' of data by companies, by requiring a user's explicit consent upon collecting personal data, or one of the other five legal bases for processing data (defined in Article 6, Lawfulness of processing (European Parliament, 2016). However, this explicit consent is gained through the notice and choice framework; a framework through which the user of a website or service clicks a button or checks a box to consent to reading the privacy statement or terms of usage and agree to the terms which the company lay-out in these statements. This framework, however, has been criticized by many scholars for it's ineffectiveness of actually notifying the user and the unrealistic expectation of each user actually reading the privacy policy (McDonald & Cranor, 2008; Cranor, 2012; Bakos, Marotta-Wurgler, & Trossen, 2009; Reidenberg, D, & Callen, 2014). This supports Zuboff's (2015) argument that the personal data is actually claimed and not provided through consent by the user.

1.1 Research

1.1.1 Problem

An information asymmetry still exists between companies and consumers regarding the usage of their (personal) data (although shortly touched upon in the introduction, this will be further elaborated on in section 2). Because of this, large tech-companies are still able to collect massive amounts of data from their users for profit, without their users knowing. As current regulations are insufficient in fully informing all users, research needs to be done towards how the problem of information inequality can be bridged. This research aims to contribute to solving this problem of inequality. In order to do this, this research will consist of exploratory analysis aimed at finding out if information can be extracted from a privacy policy which could form a signal towards indicating a company's commitment to privacy, or forms of surveillance capitalism. If successful, this method can aid in identifying which companies have a higher risk of using data in ways that do not inform their users. This knowledge can be used to get closer to a form of regulation which can control this and give more power to consumers.

1.1.2 Goal

Companies who make money off of profiling their customers or other questionable data practices often try to hide these practices. However, information on how the companies handle data and data-privacy can be derived from privacy statements, both directly or indirectly. This research will assess what information can be extracted using automated algorithms (based on natural language processing), and what can be learned from this information. These results can potentially be useful for policymakers to assess which companies need to be regulated and how, but also for DPA's whose task it is to monitor companies with a high risk to privacy. Lastly, the results can promote the writing of improved privacy statements for companies, by recognizing when a privacy policy is insufficient, be it directly from this research or indirectly via regulators. Scientifically, this research forms an extra addition to the literature aimed at automatically deriving information from privacy policies. The uniqueness of this research comes from it's focus to form conclusions on properties of the companies, in contrary to improving the readability for the users of the service.

1.1.3 Questions

Based on the research problem and goal, the following research question is defined:

Can a company's commitment to privacy be derived from their website's privacy statement using Natural Language Processing? If so, can other underlying attributes of privacy commitment be derived which could indicate surveillance capitalism? What implications does this have for privacy regulators and governments?

To answer this question, a set of subquestions is formulated which ultimately build up to answer the research question above.

Question 1 *What are privacy statements, what do they contain, and on what aspects do they differ?*

To form a starting point for the research, a literature review is performed to assess the literature on privacy statements and privacy statement analysis. From this initial research, a theoretical basis is formed in order to further build the analysis. The review should provide a clear indication of why privacy statements exist, what their purpose is, and what differentiates them. Knowing all aspects of a privacy statements makes it possible to extract the right information for the right reasons, and not extract information which may potentially be useless.

Question 2 *What variables can be extracted from privacy statements, using Natural Language Processing? What variables can be extracted within the scope of this research?*

Using the results from the previous question, an overview is created of the separate aspects through which privacy statements can differ. The variables are then attempted to be recreated and extracted from a new privacy statement dataset, which will be retrieved for this research. Each variable will be checked for accuracy, based on the achievable accuracy within the scope of this research the variable will either be used or not.

Question 3 *What information can be extracted from the variables and the privacy statement corpus? Are the results of the analyses in line with existing literature and expectations? Can information be extracted which signals privacy commitment, or surveillance capitalism?*

For this question, a large amount of analyses will be performed in chapter 5. These analyses will partly function as validation by cross-referencing known literature, and partly an attempt to identify signals of commitment to privacy and surveillance capitalism. Analyses with potentially interesting findings are presented in this thesis.

Question 4 *Do the analyses lead to results with implications for companies and regulators? What can be learned from these implications?*

The final research subquestion assesses what findings are made in the previous research questions, what implications they have on companies and regulators, and if these implications lead to important policy recommendations for policy makers. Due to the exploratory nature of this research, an emphasis will also be put on clearly defining future work.

1.1.4 Research Methods

As mentioned in section 1.1.3, the research question will be answered using exploratory, quantitative analysis, by applying the working hypothesis model.

Exploratory Analysis

When applying the method of exploratory analysis, the goal is not necessarily to find conclusive evidence to prove certain hypotheses, but to gain a further understanding of a problem or a possibility (Catterall, 2000). It is done with the goal of improving research design for future work, by developing operational definitions and establishing priorities. (Shields, 2003). The work is mostly exploratory as it is designed to be a first step in extracting commitment to privacy from a privacy statement. In a number of cases this work is guided through the conceptual framework of formulating and testing working hypotheses. These hypotheses are based on expectations and hunches rather than proved theory, and upon proving can steer further, more specified research (Shields, 2003). In this case, the problem that requires further understanding is that of finding out what companies do with data, and related to this how companies prioritize privacy. In many cases this is not evident, and few external indicators exist to assess if a company handles data in a responsible manner, or decides to sell this data without clearly notifying the user (as explained in Shultz et al.'s (2016) research). This research aims to provide a step in automatically assessing the risk of a company performing such behavior, by analyzing a company's privacy statement.

Quantitative Analysis

Analyzing these privacy statements is done by extracting variables from the statements using Natural Language Processing (section 4), and subsequently analyzing the retrieved dataset using statistical methods and clustering techniques. Based on the insights gained from these analyses and known theory, conclusions can be made on the the case of using privacy policies to gain a better understanding of the way a company deals with privacy.

1.1.5 EPA thesis

This thesis is written for the Engineering and Policy Analysis program of the TU-Delft, and thus it has to meet a number of requirements. In principle, an EPA thesis is written to contribute towards tackling a so-called 'grand challenge', and deliver policy recommendations which can help policy-makers in deciding what to do. In this case, the grand challenge that is being tackled is that of privacy, or more specifically data privacy. There currently exist challenges towards regulating companies in such a way that privacy is safeguarded for citizens, which will be further discussed in chapter 2. The current policy arrangement still allows for massive data collection and -sharing, often not to the knowledge of the consumer. Because of this, new research is needed in order to find new methods of controlling this behavior. This research aims to form a step in a new direction for a new theory of how to protect users privacy, which can ultimately (if not within the scope of this research, possibly in future work) can lead to useful policy recommendations.

Chapter 2

Literature Review

In this literature review, firstly some background is given on privacy, privacy regulations like the General Data Protection Regulation (GDPR) and privacy statements. Subsequently a short explanation of the concept of surveillance capitalism is given, as it forms an important role in the motivation for the research. Next, a short general overview is given of Natural Language Processing (NLP), and what NLP techniques can extract from natural language. Lastly, an overview is given of research where the fields of privacy statements and NLP are combined.

2.1 Privacy

Privacy was first introduced by Samuel Warden en Louis Brandeis (1890), identifying the need for laws that protect ones private life, or as they defined it "the right to be left alone". They described that there was an aspect of ones property that was not currently defended by law; it could not be described as physical or intellectual property, but consisted of information that could harm an individual, directly or indirectly, if released. An example they use is a man's collection of gems; if the man chooses to keep them secret, it is his right that others do not share that information, as sharing that information could put his household at risk of burglary. Sixty years later the EU adopted the right to privacy in it's formation under article 8, describing that "Everyone has the right to respect for his private and family life, his home and his correspondence" (Council of Europe, 1950). Article 8 is still valid, and data privacy laws as the recent GDPR (European Parliament, 2016) and the future ePR (European Commission, 2017) are partly based on this fundamental right. In this paper we therefore accept the European citizens right to privacy.

Privacy is becoming of growing importance due to it's ever-changing role in society. Technological innovations have provided the possibility to gather, process and share huge amounts of personal data in a growing number of ways. This started when Warren and Brandeis (1890) recognized the danger of sharing personal photographs of people, to now recognizing the privacy issues of a new smart thermostat (Hernandez et al., 2014). In the status quo, once a company collects data it is automatically subject to a large number of data privacy risks which multiply as the collection of data is also multiplied (Tene & Polonetsky, 2012). Because of this, data collection now comes hand-in-hand with efforts to comply with regulations like the GDPR in order to reduce privacy risks. But even compliance cannot guarantee full safety from privacy risks (Narayanan & Shmatikov, 2008; Tene & Polonetsky, 2012), meaning that it is the continuing responsibility of the company to safeguard the privacy of it's customers or users.

2.1.1 Regulations

Since the recognition for a need for privacy for civilians, regulations have been created in order to encompass this need. This need grew in the 1960s, when the rise of use of technology formed a risk to the violation of privacy (Nissenbaum, 2012). In the United States, this need for privacy became of role of the Federal Trade Comission (FTC), which has the official task of enforcing laws to prevent the violation of privacy. In the 1960s, the FTC enforced the Fair Information Practice Principeles (FIPPs) as a first privacy law in the United States (Cate, 2010). In 1980 the Organization for Economic Cooperation and Development (OECD) followed for Europe, with their privacy guidelines. After this, many new interactions and regulations followed (see figure 2.1).

GDPR

The most recent major regulation is the General Data Protection Regulation, or GDPR. The GDPR has ensured data processors (natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller (European Parliament, 2016)) to be more transparent and informative about their data collection and processing practices. Part of this is that the GDPR has set rules to ensure that certain practices of data-handling are stated in privacy statements to inform the users of what happens to their data. The party which collects, processes or shares the data is referred to as the data controller in the GDPR (European Parliament, 2016), which is defined as a party "which, alone, or jointly with other, determines the purposes and means of processing of personal data".

Under GDPR, the privacy statement must contain what personal information is collected, how it is collected, what it is used for, whether it is shared with third parties and what security measures are taken to safeguard this data (European Union, 2016). Also, it must be explicitly specified if the user has control over any of these matters. This must all be done in as little words as possible, and taking the principle of transparency into account, which states that "any information addressed to the public or to the data subject be concise, easy to understand, and that clear and plain language...be used" (European Union, 2016). These regulations aim to provide more transparency and structure to the privacy statements; something that has been advocated by scholars in the past (Cranor, 2012). The regulating authority for the GDPR is the data protection authority (DPA) (European Union, 2016), a national governmental institute which has been granted the authority to regulate and (depending on the country) often fine companies who do not abide to the rules of the GDPR.

One of the main points the GDPR makes is that the processing of personal data is heavily restricted. The GDPR defines processing as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction" (European Union, 2016). There are six legal basis for the processing of data;

- **Unambiguous Consent:** The data subject has given clear, unambiguous (and not implied) consent to the processing of their data.
- **Performance of contract:** The processing of the data is necessary in order to fulfil a contract to which the data subject has agreed
- **Legal Obligations:** The processing is necessary for the data processor to comply with legal obligations (i.e. if authorities demand the processing).
- **Vital Interest:** The processing is necessary to mitigate vital risk to the data subject or another natural person.
- **Public Interest:** The processing is to perform a task which is in the public interest that is set out in law.

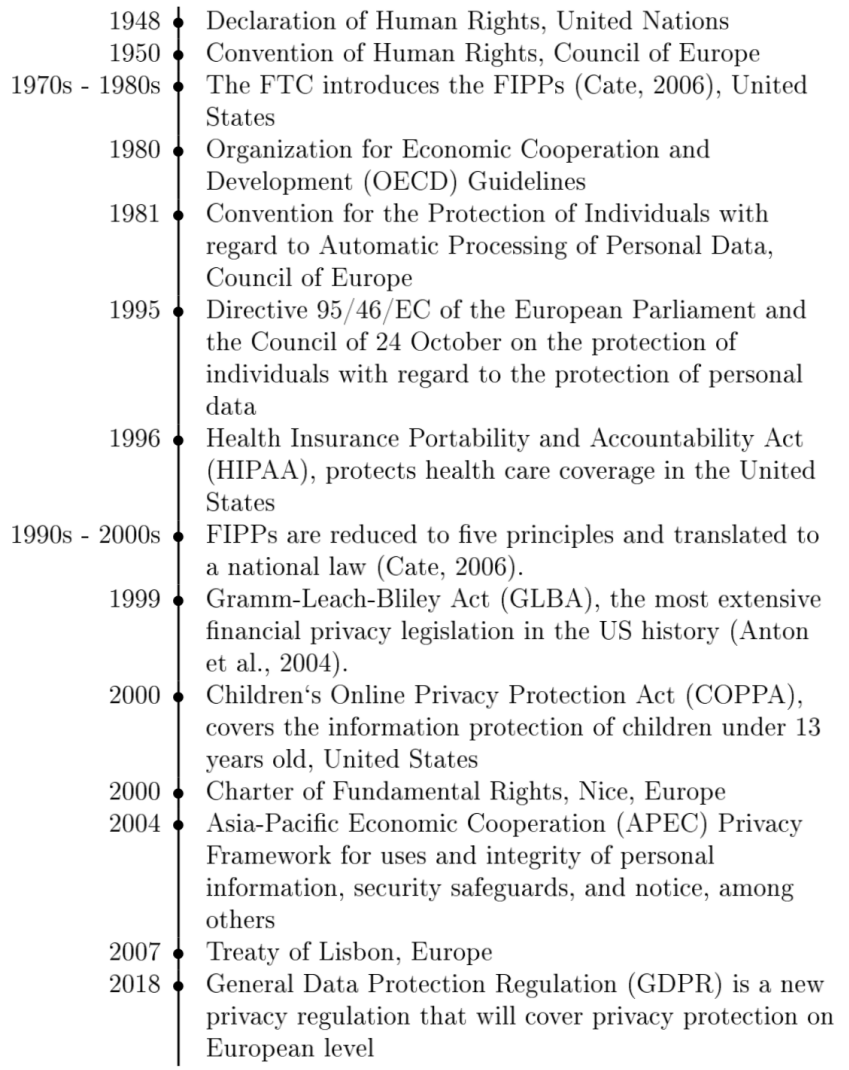


Figure 2.1: Timeline of Major US and EU Privacy Regulations (Brakenhoff, 2017).

- **Legitimate Interest:** The processing is in the legitimate interest of the user (i.e., an obvious consequence of the interest of the user).

As a data processor it is mandated by the GDPR that when data is processed, the legal basis for processed must be communicated to the user. This communication is mostly done via the privacy statement.

2.1.2 Privacy Statements

In this research, a privacy statement[1] is defined as a body of text written by a party to inform users of their services or products in which ways data is gathered, used, shared and managed. Its contents is mostly defined by the applicable regional regulations (as described in section 2.1.1), when for instance stating what personal data is collected, or legally motivated reasons, by defining a company's responsibilities in regards to the processing of data

Wilson et al. (2016) define a privacy statement annotation scheme in their research in analyzing privacy policies from a large corpus of websites. This annotation scheme was created by domain experts in the field of privacy and public policy, and where created to capture the main topics a complete privacy statement should cover. The eventual ten data practice categories of the annotation scheme are the following:

- *First Party Collection/Use*: what types of user data is collected and used and why
- *Third Party Collection/Use*: how user information is shared and why
- *User Choice/Control*: available choices and control for users
- *User Access, Edit, —& Deletion*: if and how users can access, edit and delete their information
- *Data Retention*: how long information is stored
- *Data Security*: how user information is protected
- *Policy Change*: if and how users will be informed about changes to the privacy policy
- *Do Not Track*: if and how Do Not Track signals are mentioned
- *International & Specific Audiences*: if specific groups of users are mentioned
- *Other* additional sub-labels for parts of the text without content (i.e. title, contact information)

These data practice categories were defined before the GDPR was announced; as such they form a useful evaluation metric outside of GDPR requirements to assess the completeness of a privacy statement. This is useful as the GDPR had a profound effect on the way privacy statements are written (Libert, 2018).

The privacy statement covers these topics, with the goal to inform the user of how and why data is collected. Under GDPR the data controller is obliged to notify the data subject when they process personal data, on what legal basis this is done. The notice from the company is often in the form of a privacy statement. Implying that the user reads the company's privacy statement, he or she then has the power to give informed consent, essentially controlling what happens to his or her personal information. This framework has formed the basis of information privacy regulations in both the United States and European Union (Schwartz & Solove, 2009), including the GDPR and ePR.

While notice and choice relies on every user reading and understanding the notice by the company, McDonald and Cranor (2008) calculated that if every US citizen read the notices they would need 244 hours per year.

[1]This research explicitly chooses to use the term "privacy statement". In some cases companies also use the terms "privacy policy" or "privacy notice" to reference the same document. However, these words may lead to confusion, as for instance a company's internal policy towards privacy can be referred to as a privacy policy, and (more informally) governmental regulations concerning privacy are referred to as privacy policies. Likewise, when a user of a product or service receives a notification of a matter that regards the user's privacy, this may be referred to as a "privacy notice". To reduce ambiguity, only the term privacy statement will be used, which is defined in the section above.

With the growing connectivity of devices and the therefore increasing number of privacy statements that need to be read, this number has probably enlarged compared to when this research was performed in 2008. Furthermore these policies are often too complex, limiting users' understanding of the choices they are given (Cranor, 2012). This contributes to these notices failing to inform their users and largely being ignored (Bakos et al., 2009; Cate, 2010). Reidenberg et al. (2014) have further criticized the notice and choice framework, stating that even if fully read and understood by the user, the ideal working framework only solves part of the problem; wrongful retention and security of the data are still reliant on the company.

2.1.3 Surveillance Capitalism

A new concept within data privacy which is of importance for this research is *surveillance capitalism*. This section gives a brief overview of what surveillance capitalism is, and why it is important for this research.

The term surveillance capitalism as will be used in this research was first introduced by Shoshanna Zuboff (2015) and explains the economic logic behind why companies are gathering data. As she describes, surveillance capitalism is a "new form of information capitalism [which] aims to predict and modify human behavior as a means to produce revenue and market control." She elaborates on this by describing that a fourth fiction has been added to the three fictions for which market economics were created, as described by Karl Polanyi in "The Great Transformation" (1978). The three fictions Polanyi described are the redistributive fiction (the predecessor of the labour market), the reciprocity fiction (the predecessor of currencies) and the householding fiction (predecessor of the real estate market).

This fourth fiction according to Zuboff is the behavioral fiction; due to the evolution of technology, reality is now measurable and subsequently commodifiable. Zuboff continues on this subject in her book (Zuboff, 2019), where she further highlights that the data, after being gathered and processed, are traded in so called *behavioral future markets*. Here, after being gathered and processed by the surveillance capitalists, the data on the users' futures are being bought by advertisers, insurers, retailers, companies in finance, and an ever-growing group of goods- and services companies. According to Zuboff, this data is freely and unilaterally acquired with notice and choice systems. This keeps an information asymmetry in place between the surveillance capitalist and the user of the service, implying that the user is unknowingly sharing his data and unaware of the consequences that might follow. When combining this with the earlier definition of privacy by Altman (1975), the privacy issues become apparent. In the situation Zuboff describes, the individual is unable to distinguish what data is private and what data is public anymore. In a time where the United Nations have declared that the internet is a basic human right (Human Rights Council, 2016), citizens are dependent on the services that the surveillance capitalists provide (Zuboff, 2015).

Although Zuboff does not provide a single definition of a surveillance capitalist, her explanations on aspects of surveillance capitalism provide clear descriptions of what practices are common to surveillance capitalists;

> "Surveillance capitalism's products and services are not the objects of a value exchange. They do not establish constructive producer-consumer reciprocities. Instead, they are the "hooks" that lure users into their extractive operations in which our personal experiences are scraped and packaged as the means to others' ends. We are not surveillance capitalism's "customers." Although the saying tells us "If it's free, then you are the product," that is also incorrect. We are the sources of surveillance capitalism's crucial surplus: the objects of a technologically advanced and increasingly inescapable raw-material-extraction operation. Surveillance capitalism's actual customers are the enterprises that trade in its markets for future behavior." (Zuboff, 2015)

From this quote can be understood that a surveillance capitalist can be recognized by it's business model; while the company produces products for consumers, these are (at least partly) created to extract data from the consumers. This data then is used to create predictions of future behavior of the consumers, which are the actual products that provide income. The best example of a surveillance capitalist is Google, which according to Zuboff is the inventor of surveillance capitalism. Google's search engine is offered as a free

service because it is used as a tool to collect personal data, and providing the service for free maximizes the amount of users. This data then functions as the raw material for their actual business model, which is selling the prediction products of their users to third parties (Esteve, 2017; Zuboff, 2019).

In another quote, Zuboff describes the distinction between capitalism and surveillance capitalism:

> "...[the] line [between surveillance capitalism and capitalism] is defined in part by the purposes and methods of data collection. When a firm collects behavioral data with permission and solely as a means to product or service improvement, it is committing capitalism but not surveillance capitalism." (Zuboff, 2015)

An identifier of a surveillance capitalist is therefore their intent with data, meaning a distinction can be able to be made between data usage which indicates surveillance capitalism and data usage which does not. The factors explored here by Zuboff are of importance to this research. The way a company handles data has an important role in if a company is a surveillance capitalist, and therefore in a company's attitude towards the privacy of it's users. As described in section 2.1.2 and section 2.1.1, the privacy statement of a company contains important aspects of the way a company handles information, as it is required by law.

2.1.4 Commitment to Privacy

Although surveillance capitalists have a natural link to privacy as their business model directly influences the privacy of the users, most companies do not necessarily have data privacy as a high priority. Some companies do not process any personal data at all, making it natural that these companies have privacy as a low priority. As the relevance of having a well thought-out policy for privacy can vary, but that the effects of privacy can have an indirect positive effect on society, maintaining privacy is increasingly seen as a corporate social responsibility (CSR) (Pollach, 2011; Garriga & Melé, 2013). Corporate social responsibilities concern societal responsibilities which are not necessarily in the direct interest of businesses as they do not directly contribute to the success of the business. Privacy is now often seen as a corporate social responsibility as it also improves the well-being of society, in the form of users feeling that their data is being used responsibly (Pollach, 2011). This further emphasizes one of the issues of privacy, that it is (often) not in the direct interest of the company, as putting time and effort into an internal privacy policy is not directly rewarding.

In her research, Pollach (2011) assesses online documents from a large number of IT-companies to find out how they handle privacy. The goal of this research is to assess the motives behind a company's commitment to privacy, which is defined as the degree to which a company sees privacy as a CSR, and correctly protects the privacy of it's users or customers. She decides to only analyze documents directly explaining CSR policies as they are the most explicit. The level of commitment of the company can be attributed to a number of factors, which Pollach recognized as moral, relational and instrumental motives (Pollach, 2011). Moral motives have no direct relation to the company, but concern the responsibilities the company feels it has towards its users (for instance simply acknowledging that their customers have a right to privacy). Relational and instrumental motives regard arguments which indirectly benefit the company, like winning customer trust and gaining a reputational advantage. Lastly, she recognizes that many companies that have a lesser priority for privacy do so as they lack the resources, indicating their time is more valuable to spend in other business-related matters. Furthermore, the research also contains the types of actions which are performed by the IT companies to safeguard privacy. Here it was found that in every case at least the privacy statement was present, which makes sense as the privacy statement is mandatory by law. Because of this, the privacy statement always forms an indication to privacy, to some degree. To what degree exactly is not part of the research.

As described in the previous subsection (section 2.1.3), Zuboff (2019) also recognizes that business models and the usage of data influence how companies handle data privacy, partially directly related to surveillance capitalism, and partially due to the fact that a company's policy towards privacy is simply influenced by the way a company uses data. These two works combined therefore provide an indication of what influences

the way a company forms their corporate social responsibility of privacy (or more simply, commitment to privacy).

2.2 Natural Language Processing

Natural language processing (NLP) is the process of manipulating natural languages like English or Dutch with computers, in order to extract information related to the context of that document (Bird, Klein, & Loper, 2009). By doing this, information can automatically be extracted from the document without having to manually read it, assuming the correct NLP-techniques are applied. There are a large number of of techniques related to NLP, which can all be used in separate ways. Because of this, the information extraction possibilities related to NLP and the possible ways in using these methods are abundant (Bird et al., 2009).

Falessi et al. (2013) have created a framework of available NLP techniques in their research, which provides a suitable overview of techniques for this literature review. This overview can be seen in figure 2.2 where each of the four 'variation points', as Falessi et al. call them, can be seen as axes upon which an NLP-technique can vary. These variation points are algebraic models, term extraction, weighting schemas and similarity metrics. They also acknowledge that the model does not cover every single aspect of NLP, as it does not cover word sense disambiguation, Latent Dirichlet Allocation (LDA), Relational Topic Models and probabilistic IR techniques. Although LDA-modelling was tested but ultimately not used in the final research, the framework proves to be a suitable overview for this research.

2.2.1 Techniques

In this subsection, each of the four categories is shortly discussed.

Term Extraction

Falessi et al. (2013) describe the pre-processing of text and first syntactic analysis as term extraction, which they divide into two groups, namely *simple* and *POS-tagging*. The goal of these steps is to prepare the text for analysis; unprocessed written text has less concisive information which can be extracted from it than pre-processed text (Bird et al., 2009).

Simple can for instance consist of identifying each word or sentence (tokenizing, see figure 2.3), removing punctuation and capitals, removing spelling-errors and removing stop-words (such as a, the, as, to etc.) (Falessi et al., 2013). These steps are necessary to provide the most accurate results after applying models to the processed texts. For instance, if a certain construction only exists within sentences, separating each sentence and analyzing it makes it possible to extract this construction.

A next step in the process can be **Part-of-speech tagging** (POS-tagging). POS-tagging entails tagging the grammatical function of each word in a sentence, e.g. a noun, verb, adjective etc. (see figure 2.5). This provides the possibility for the algorithms of gaining a better understanding of the build up of sentence. Falessi et al. (2013) describe two different ways of using the tags, namely for term-weighting and stemming. The first part entails labelling each word according to it's "expected amount of semantic contribution in the given application domain" (Falessi et al., 2013), or in other words applying a value to the meaning of a text. For instance, sentences which share the same nouns will have a higher similarity than texts which share the same articles. The second part, stemming consists of reducing each word to its grammatical base (i.e. translating the words "monitoring", "monitor", "monitored" and "monitors" into the word monitor. An example of a stemmed sentence can be seen in figure 2.4. Stemming functions as a method for the computer

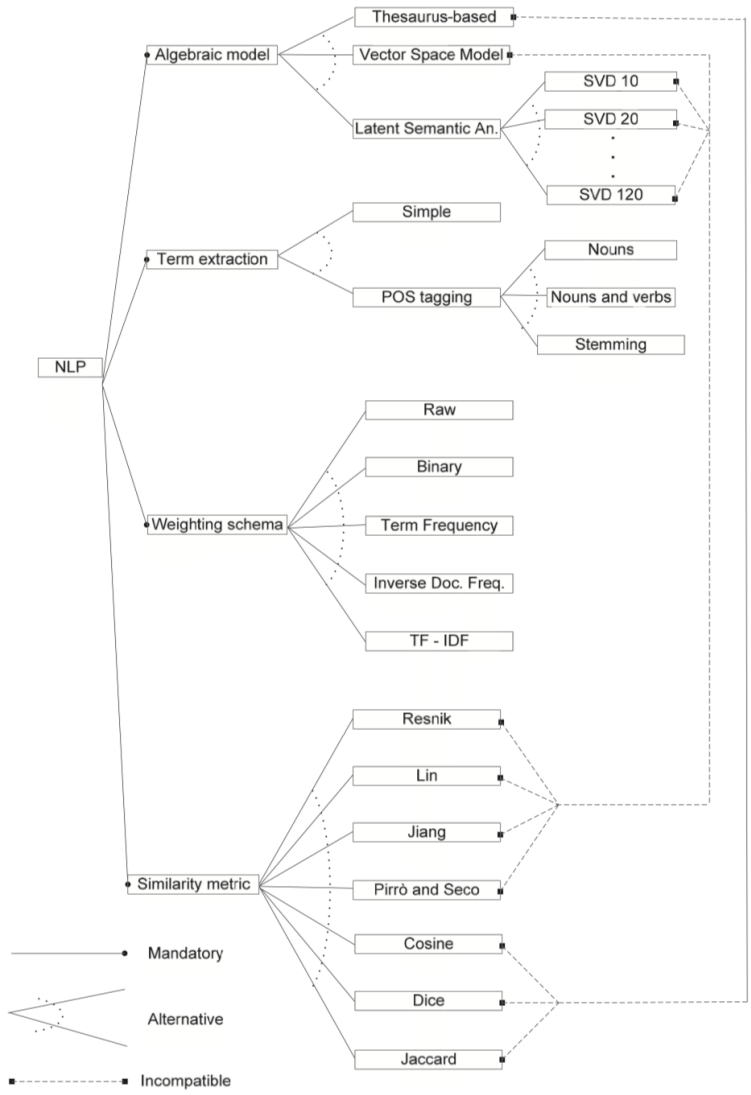


Figure 2.2: Overview of NLP techniques, from Falessi et al. (2013)

```
[AWeber, Systems, Inc., is, responsible, for, the, processing, of, personal, data, it, receives, under, each, Privacy
, Shield, Framework, and, subsequently, transfers, to, a, third, party, acting, as, an, agent, on, its, behalf, AWebe
r, Systems, Inc., complies, with, the, Privacy, Shield, Principles, for, all, onward, transfers, of, personal, data,
from, the, EU, the, United, Kingdom, and, Switzerland, including, the, onward, transfer, liability, provisions]
```

Figure 2.3: Tokenized sentence, from the privacy statement of *Aweber Systems*.

```
['AWeber', 'Systems', 'Inc.', 'responsible', 'processing', 'personal', 'datum', 'receive', 'Privacy', 'Shield', 'Fram
ework', 'subsequently', 'transfer', 'party', 'act', 'agent', 'behalf', 'AWeber', 'Systems', 'Inc.', 'complie', 'Priva
cy', 'Shield', 'Principles', 'onward', 'transfer', 'personal', 'datum', 'EU', 'United', 'Kingdom', 'Switzerland', 'in
clude', 'onward', 'transfer', 'liability', 'provision']
```

Figure 2.4: Stemmed paragraph from the privacy statement of *Aweber Systems*, without stop words

to gain a fuller understanding of each sentence; once each word is grammatically labeled, the correct base of the word can more easily be identified, and these bases give a clearer indication of the meaning of a text.

```
[['AWeber', 'PROPN'], ['Systems', 'PROPN'], ['Inc.', 'PROPN'], ['is', 'VERB'], ['responsible', 'ADJ'], ['for', 'ADP']
, ['the', 'DET'], ['processing', 'NOUN'], ['of', 'ADP'], ['personal', 'ADJ'], ['data', 'NOUN'], ['it', 'PRON'], ['rec
eives', 'VERB'], ['under', 'ADP'], ['each', 'DET'], ['Privacy', 'PROPN'], ['Shield', 'PROPN'], ['Framework', 'PROPN']
, ['and', 'CCONJ'], ['subsequently', 'ADV'], ['transfers', 'NOUN'], ['to', 'ADP'], ['a', 'DET'], ['third', 'ADJ'], ['
party', 'NOUN'], ['acting', 'VERB'], ['as', 'ADP'], ['an', 'DET'], ['agent', 'NOUN'], ['on', 'ADP'], ['its', 'DET'],
['behalf', 'NOUN'], ['AWeber', 'PROPN'], ['Systems', 'PROPN'], ['Inc.', 'PROPN'], ['complies', 'NOUN'], ['with', 'ADP
'], ['the', 'DET'], ['Privacy', 'PROPN'], ['Shield', 'PROPN'], ['Principles', 'PROPN'], ['for', 'ADP'], ['all', 'DET'
], ['onward', 'ADJ'], ['transfers', 'NOUN'], ['of', 'ADP'], ['personal', 'ADJ'], ['data', 'NOUN'], ['from', 'ADP'], ['
the', 'DET'], ['EU', 'PROPN'], ['the', 'DET'], ['United', 'PROPN'], ['Kingdom', 'PROPN'], ['and', 'CCONJ'], ['Switze
rland', 'PROPN'], ['including', 'VERB'], ['the', 'DET'], ['onward', 'ADJ'], ['transfer', 'NOUN'], ['liability', 'NOUN
'], ['provisions', 'NOUN']]
```

Figure 2.5: POS-tagged sentence from the the privacy statement of *Aweber Systems*

Similarity Metrics

Similarity metrics concern formulas which in different methods compute the similarities between two texts. As can be seen in figure 2.2.1, Falessi et al.(2013) describe 7 similarity metrics, of which 4 are *WordNet* metrics. The three 'normal' metrics measure to what degree the same words are used in different texts, without taking semantics into account. The WordNet similarity metrics however are based on a large pre-trained lexical database (Fellbaum, 2005). For instance, when analyzing a piece of text using the WordNet database, analyzing the word "Barack Obama" would result in the lexicon recognizing this is a president, as opposed to an untrained algorithm which would simply recognize this as a name. This provides the possibility of creating more accurate similarity metrics.

Weighting Schema

Weighting schemes concern multiple techniques that assign values to words in a text, based on the frequency the word occurs in the text. Using these metrics allows the possibility of translating the presence of a word into numerical values, which can be used for for instance comparisons of word-usage between section of a document or between multiple documents. Falessi's (2013) describes five different types of term weighting methods, namely:

- Raw Frequency: The number of occurrences of a specific word.
- Binary Frequency: Assesses the presence of a word with a binary value.
- Term Frequency: calculates the frequency of occurrences of the word relative to the length of the text. Useful to understand how often certain words are used.
- Inverse Document Frequency: assigns a value based on the number of given text fragments which contain a specific word. Useful to understand which sentences contain a specific word.

- TF_IDF: Multiplication of the two metrics above, which provides the possibility of using both metrics for one term, and giving insight into both uses in one metric.

The first two metrics described by Falessi et al. (2013) are useful for quickly assessing the presence of a certain word (i.e. a certain topic) in a text, while the latter three give an insight into how a certain word is used in a text. The formulas of the latter three measures are shown in figure 2.6

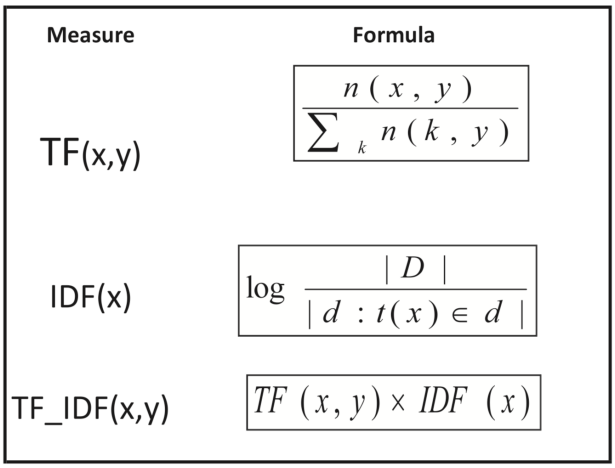


Figure 2.6: Term Weighting Formulas, from (Falessi et al., 2013)

Where for Term Frequency, the sum of the occurence of word x in document y (shown in the formula as $n(x, y)$) is divided by total occurences of all words k in document y (shown in the formula as $\Sigma n(k, y)$).

For Inverse Document Frequency, the value is log-value of the total number of documents $| D |$ divided by the total number of documents where term *t(x)* occurs, i.e. $| d : t(x) \in d |$.

Lastsly, the TF_IDF is calculated by multiplying both formulas, when measuring one value for x (i.e. for one word).

Algebraic models

Algebraic models are used to understand similarity in meaning between words, or semantic similarity. Falessi et al. (2013) divide algebraic models into three groups, namely thesaurus-based models, vector space models (VSM) and latent semantic analysis (LSA). Using these techniques, documents of words can be translated into vectors, which in turn give insights into similarities in the usage of words within a document. This can be done without a thesaurus (VSM) or with a thesaurus (thesaurus-based, LSA). Benefits of using a thesaurus when vectoring text is the possibility to have synonyms pre-defined. For instance, where VSM would not know the similarities between the terms 'car' and 'vehicle', thesaurus-based techniques would be able to extract the coefficient of similarity for these two terms.

2.2.2 Readability Metrics

Not part of natural language processing, but also relevant for understanding aspects of written language are readability metrics. Readability metrics are metrics resulting from formula's which aim to calculate the level of complexity that at text exhibits when reading. The most well-known readability metrics for the English language are the Flesch-Index (Flesch, 1949), the Dale-Chall index (Dale & Chall, 1949), the SMOG-index

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Figure 2.7: Flesch Reading-Ease (FRE) Formula (Flesch, 1949)

(McLaughlin, 1969), and the Fleisch-Kincaid formula (Kincaid, Fishburne, Robert P., Richard L., & Brad S., 1975). Even though multiple readability metrics exist, for consistency it is recommended to only use a single metric for analysis: different readability metrics tend to less accurately measure documents that are either very complex or very simple, meaning some metrics will be better suitable to measure complexity than others (Zamanian & Heydari, 2012).

The *Flesch Readability Ease Score* was introduced in the book "How to write plain English" by Robert Flesch (1949), and uses the following formula:

The achieve the scores, the average words per sentece and the average syllables per word are calculated. The formula results in a score between 0 and 100, where 0 indicates the text is practically unreadable and 100 indicates the text is extremely easy (Flesch, 1949). The score can also be translated to readability levels which explain the reading level, ranging from 'Very Easy' to 'Very Difficult'. The score is still widely used, and praised for being usable for complex and simple tests alike (Fitzgerald & Watkins, 2006).

FLESCH READING EASE SCORE

Reading Ease Score	Description	Predicted Reading Grade	Estimated Percentage of U.S. Adults
0-30	very difficult	college graduate	4.5%
30-40	difficult	college grade	33%
50-60	fairly difficult	10^{th}-12^{th} grade	54%
60-70	standard	8^{th}-9^{th} grade	83%
70-80	fairly easy	7^{th} grade	88%
80-90	easy	6^{th} grade	91%
90-100	very easy	5^{th} grade	93%

Figure 2.8: Interpretation of Flesch-Index, from (Zamanian & Heydari, 2012)

The Dale-Chall index (Dale & Chall, 1949) improved on the Flesch-index by adding a dictionary of difficult words to further enhance the score. This dictionary consisted of 763 words, but was later expanded to 3000 words (Kiyokawa, 1996). The index has proved to be accurate for many general texts, as it incorporates words which are difficult to understand.

The 'Simple Measure of Gobbledygook', or SMOG-index, uses three samples of sentences from the text to estimate what school-grade a reader would need to have passed in order to be able to easily read the text (McLaughlin, 1969). Although praised for it's simplicity, in some cases it can also be seen as it's drawback; because it takes samples of the text, the assumption is made that the full text exhibits the same levels of complexity throughout (Zamanian & Heydari, 2012).

Lastly, the Flesch-Kincaid formula (Kincaid et al., 1975) concerns a newer, more simplistic version of the Flesch Reading Ease formula. It was conceived for educational purposes, and results in the expected (school) grade a reader needs to have in order to be able to read the text. This formula is a variation of the Flesch-formula purely created to simplify the outcome. It is based on the same variables as the Flesch formula (sentence length and word length).

2.3 Privacy Statement Analysis

Using the techniques described in the previous section (2.2.1), many attempts have been made towards analyzing privacy statements in order to extract information for a variety of purposes. The majority of these (Costante, Sun, Petković, & den Hartog, 2012; Sadeh et al., 2013; Liu, Ramanath, Sadeh, & Smith, 2014; Mysore Sathyendra, Schaub, Wilson, & Sadeh, 2016; Wilson et al., 2016) to name a few, have been aimed at improving privacy statement readability for the user. This is for instance done by automatically extracting opt-out choices or automatically assigning certain sections of text to certain privacy statement topics. In this section, a selection of the 3 most relevant papers for this research are shortly reviewed.

2.3.1 Vagueness

The first paper reviewed is that by Bhatia et al. (2016), called 'A Theory of Vagueness and Privacy Risk Perception. The paper uses empirical content analysis to derive a thoery of vagueness for privacy statements, based on a taxonomy of vague terms derived from manual analysis of 15 privacy statements. The research is motivated by the fact that vagueness has the potential to conceal privacy-threatening practices, and decreases the possibility of users to make informed choices, which in turn may increase their perceived privacy risk. After manual annotation, they define four vagueness categories:

- Conditionality - the related action is dependant on an unclear variable
- Generalization - the related action is unclearly defined
- Modality - the likelihood of the action is ambiguous
- Numeric Quantifier - the quantifier of the action is vague

After performing content analysis, a taxonomy is provided for each vagueness category, which can be seen in table 2.1. Within this taxonomy, they found that mostly sentences containing both 'generalization' and 'modality' words contributed most to perceived vagueness.

Table 2.1: Taxonomy of vague terms for privacy statements

Category	**Vague Terms**
Conditionality (C)	depending, necessary, appropriate, inappropriate, as needed
Generalization (G)	generally, mostly, widely, general, commonly, usually, normally, typically, largely, often
Modality (M)	may, might, can, could, would, likely, possible, possibly
Numeric Quantifier (N)	certain, some, most

2.3.2 Polisis

An elaborate attempt of automated privacy statement analysis is that of Harkous et al. (2018), who have created the tool *Polisis*. This tool is a tool built with deep-learning, trained on 130k privacy statements, and is able to extract fine-grained privacy related information, which is of use to the user. This is done by separating each statement into smaller groups of texts, or segments. These segments are then labeled using labels relating to different topics which can be covered in privacy statements. The final result is their tool

Polisis, which is freely available online [2], and able to dissect each privacy statement and with fairly high accuracy display what important information can be retrieved for the user of the service.

Their top-level labeling is based on the taxonomy by Wilson et al. (2016), discussed in section 2.1.2. Upon analyzing a segment, the classifier decides to which of these categories a segment belongs, whereby multiple categories are also possible. On a more lower-level, privacy attributes are defined, and within the attribute a specific label is defined. The overview of categories, attributes and labels can be found in figure 2.9. This overview provides a clear framework of expected types of information that are contained within a privacy statement.

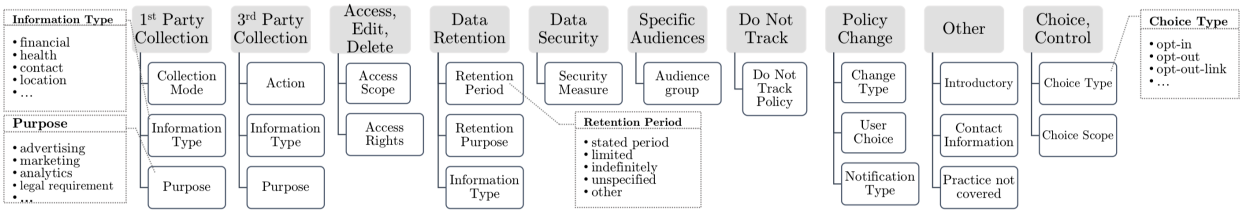


Figure 2.9: Categories, attributes and labels, used for Polisis (Harkous et al., 2018), also used as a general overview of topics used in a privacy statement.

2.3.3 The Privacy Policy Landscape After the GDPR

The second research, by Linden et al. (2018), uses the reading capabilities of the Polisis tool with a large corpus of pre-GDPR and post-GDPR privacy statements. The goal of their research is to identify what differences can be found between pre- and post-GDPR privacy statements. To do this, a number of additional variables are extracted on top of those extracted in the research by Harkous et al. (2018). Where the variables that Harkous et al. (2018) uses aim to extract specific information on what the privacy statement covers and how data handling is described, Linden et al.'s (2018) variables have a more general focus, aimed at giving a score to a certain aspect of the privacy statement. These are the following:

- Readability: Assessing how easily readable statements are. This is not done by using conventional readability metrics as described in section 2.2.2, but assessing average number of syllables, words, sentences, words per sentence, and number of sentences containing passive language.
- Presentation: The lay-out of a privacy statement. This is assed manually, asking 470 participants to rate the clarity of the statements.
- Coverage: Assesses the number of topics covered, based on the topics define in the research of (Wilson et al., 2016). See figure 2.9
- GDPR-Compliance: Assessed by using a checklist from the UK's Information Commissioner's officer's guide to GDPR [3]. Answers to the checklist items are retrieved using the Polisis tool.
- Specificity: uses the Polisis framework to assess the average level of specificity of the privacy statement. This is extracted by assessing the granularity of the information Polisis retrieves, which indicates to what degree a privacy statement explicitly defines their data practices.

Linden et al. found that although coverage and GDPR-compliance improved, this was at cost of the length, readability and sometimes specificity of the privacy statements. This means that privacy statements contain more information but are harder to read, further emphasizing the potential of automatically extracting

[2] www.pribot.org/polisis
[3] https://ico.org.uk/for-organisations/guide-to-the-general-data- protection-regulation-gdpr/

information; algorithms do not suffer from decreased readability, but are able to extract the additional information.

The research described in the previous sections (2.3.2 and 2.3.3) are two very recent and elaborate attempts at extracting a broad selection of variables, and thus useful information from privacy statements. An overview is given of the variables from both papers in table 2.2.

Table 2.2: Privacy Statement Variables

Variable Group	**Description**	**Source**
Length	The amount of words used in the full privacy policy, or lexicon count.	(Linden et al., 2018)
Readability	Index for the ease of reading of a document, for instance the Flesch Kincaid Grade (Kincaid et al., 1981) or the Dale-Chall Raedability Score (Dale & Chall, 1949)	(Linden et al., 2018)
Presentation	Manual assessment of the overview of a document, not based on the content but more on the "look and feel" of the document.	(Linden et al., 2018)
Coverage	Coverage concerns if the privacy policy covers a number of pre-defined privacy categories, such as First party collection, Third party collection, User Control and more.	(Linden et al., 2018)
Compliance	Compliance is a more specified version of coverage and looks at a specific set of 7 questions regarding compliance with the GDPR.	(Linden et al., 2018)
Specificity	Specificity concerns to what degree specific information is provided in favor of terms being kept vague. For instance, "We collect your personal information..." is less specific than "We collect your health data...".	(Linden et al., 2018)
Types of information collected	Types of information collected concern a number of binary variables, each portraying a type of data which can be collected by the website, for instance computer information, contact information, demographic infomration, location etc.	(Harkous et al., 2018)
Collection reasons	Collection reasons concern a number of binary variables, each portraying the reason why a website collects data, for instance Analytics Research, Additional Service Feature, Basic Service Feature etc.	(Harkous et al., 2018)
Types of information shared	Types of information shared consists of the same categories as Types of information collected, only now the binary variables represent if the data is also shared with third parties.	(Harkous et al., 2018)
Sharing reasons	Sharing reasons consists of the same categories as collection reasons, only now the binary variables represent why the data is shared with third parties.	(Harkous et al., 2018)
Opt in/out choices	The choices the privacy policy present over which the user can control their data	(Harkous et al., 2018)
Security	Assessment of how the privacy policy handles different aspects of security.	(Harkous et al., 2018)
Data retention	What data does the company retain and how long?	(Harkous et al., 2018)
Specific audiences	Does the policy discuss specific audiences or specific places?	(Harkous et al., 2018)
Rights to edit	Does the company give you the option to edit their data?	(Harkous et al., 2018)
Policy change	What happens when the company changes its policy?	(Harkous et al., 2018)

2.4 Conceptual Framework

To bring all this material together and to guide the research, a conceptual framework is presented in this section. The conceptual framework in this research serves as a schematic overview of what has been covered so far and what is proven in literature, and an indication of where this research will attempt to provide new insights. See figure 2.10.

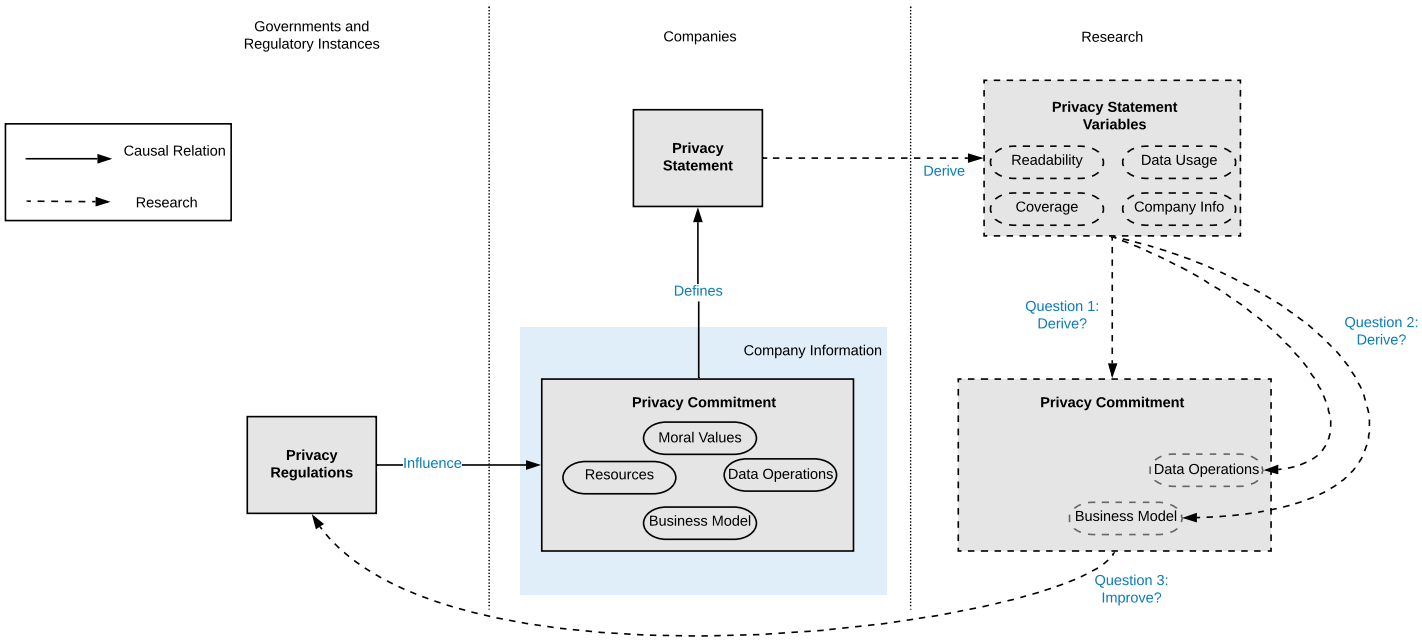


Figure 2.10: Conceptual Framework: Read from left to right. Full arrows mark relations which have been proved in previous research, dotted arrows mark causal relations this research will test

Firstly, each part of the conceptual framework is linked to the handled literature in this section. Starting left in the conceptual diagram, the influence of regulations on a company's commitment to privacy is indicated, which is explained in section 2.1.1. Moving to the right to companies, privacy commitment, which is explained in section 2.1.4, is influenced by a number of factors related to the company, treated in section 2.1.4 and 2.1.3. A small side note on these factors is that in reality they are also expected to influence each other, i.e. the type of business model a company chooses will likely influence its data usage and its resources, but for clarity of the framework these relations are left out. Based on a company's commitment it feels towards privacy, it will put effort into it's internal policy for privacy, which is explained in Pollach's (2011) research, including it's privacy statement. The privacy statement, as explained in section 2.1.2, is then published on the website (or via other means) by the company, making it publicly available.

From this point the relations in this diagram turn into the relations which are to be proved in this research. Once the statements are made publicly available, they can be gathered and used for research. In this specific research, an attempt will be done to use natural language processing to derive different variables relating to the privacy statement, such as readability, data usage, coverage and company info. This is essentially creating a numerical representation of the privacy statement, which can be used for analysis. From this extracted data, three steps of research are attempted. Firstly, an attempt will be made to use this variables to derive some indication of a company's commitment to privacy. As said before, commitment to privacy is a result of multiple factors within a company. Here, it is assessed to what degree this level of commitment

can be extracted from the documents. Next, the second step focuses on extracting information related to surveillance capitalism. As described by Zuboff (2019) and explained in section 2.1.3, surveillance capitalists differentiate themselves from regular capitalists by their usage of data. Large amounts of data are collected and shared, often for marketing purposes. Because of this, variables related to data usage (like which types are collected and shared, and for what purpose) could indicate surveillance capitalism. In the second step of this research this possible link will be assessed. Finally, based on the results of the first two steps, a recap of all findings and their implications will be done to see if policy recommendations can be made which can improve future privacy regulations.

2.4.1 Research Questions

Based on the literature discussed in this chapter and the provided conceptual framework, the following problem statement is formulated:

Can a company's privacy commitment be derived from their website's privacy statement using Natural Language Processing? If so, can other underlying attributes of privacy commitment be derived which could indicate surveillance capitalism? What implications does this have for privacy regulators and governments?

The problem statement shortly describes each of the three dotted connections which are laid out in the conceptual framework. In the first question, a choice is made to only use syntactic forms of NLP. The reason for this is mostly to define the scope of this research. A large part of this research will consist of collecting a database of privacy statements, and applying NLP to these statements to extract variables.

To answer these questions, four sub questions are defined which ultimately make it possible to answer the main research question(s). These are;

Question 1 *What are privacy statements, what do they contain, and on what aspects do they differ?*

Question 2 *What variables can be extracted from privacy statements, using Natural Language Processing? What variables can be extracted within the scope of this research?*

Question 3 *What information can be extracted from the variables and the privacy statement corpus? Are the results of the analyses in line with existing literature and expectations? Can information be extracted which signals privacy commitment, or surveillance capitalism?*

Question 4 *Do the analyses lead to results with implications for companies and regulators? What can be learned from these implications?*

The first question has already been answered in the literature review. The second question will be answered in chapter 4, the third question will be answered in chapter 5, and the last question will be answered together with the presentation of the conclusions, in chapter 7.

Chapter 3

Obtaining the Dataset

This section describes how a new database of privacy policies is obtained. The process consists of the following steps:

- Website selection
- Creating an Amazon Mechanical Turk (MTurk) project
- Executing MTurk project
- Extracting test subsets
- Extracting the full database

The sections below describe how each step is performed.

3.1 Website Selection

The first step in creating the database is selecting from which websites privacy statements will be extracted, for which a number of conditions need to be taken into account. As the goal of this analysis is to extract privacy statements it is important to make a selection of websites that have a privacy statement. Another important factor to take into account is that the eventual database of privacy statements will be analyzed using Natural Language Processing (NLP). Most NLP-techniques are language specific (see section 2.2.1, meaning extracting the same information from privacy statements in differing languages would require creating (and testing) multiple algorithms, one for each language. As this costs extra time and increases possible inconsistencies in the research, the analysis is limited to privacy statements of one language, namely English.

To maximize the results, a selection is made from a list of most visited websites worldwide. This list is obtained from Majestic and is called the Majestic Million database [1]. The database is a list of 1 million websites, ranked by the number of referring subnets and referring IPs. Two main benefits of this database are that firstly it is free to download, and secondly the data also contains a field of the website's Top-Level Domain (TLD). This field is used to control for the second condition, namely that of English language. To ensure the maximum amount of English websites are analyzed, the websites are filtered on their TLD. The top-5 international or English TLDs are selected from the database, which are *.com, .net, .org, .eu and .int*. Of the remaining websites, the top 2000 most visited are kept. The amount of 2000 was chosen due to the aim of having a database of at least 1000 privacy statements, which is enough to perform multiple statistical analyses, and costs, which are implied upon using Amazon MTurk (section 3.2.

[1] https://majestic.com/reports/majestic-million

3.2 Amazon Mechanical Turk

Extracting a privacy statement from a website is not a task that can easily be automated. Due to differing formats and locations on privacy statements on websites, performing an automated crawl would likely not result in an accurate database. Some attempts have been made to perform automated webcrawls for privacy statements (Libert, 2018), but these methods have not yet been made publicly available or tested for accuracy. Therefore, for this research an Amazon Mechanical Turk (MTurk) project is created. MTurk is a crowdsourcing marketplace created to outsource processes and jobs to a distributed workforce. In other words, using MTurk simple tasks can be outsourced online to a large amount of workers against a small fee. In this case the task consists of clicking on one of the links from the selection explained in the previous section, finding the privacy policy linked to that website and copying this in a specified input.

To ensure input could be checked for validity, each website was checked by three different workers, resulting in a total of 6000 assignments. Per assignment, a worker is asked to go to a certain link (being one of the 2000 selected above), and to find the privacy statement of that website. Once found, the worker is asked to first copy the link of the privacy statement, followed by the date the privacy statement was last updated, and finally copy the full privacy statement. The worker received a fee of $0.05 (USD) for each assignment, resulting in a total cost (including a 20% fee for Amazon) of $360. Figure 3.1 shows what the task looked like for workers, which was written using *JavaScript*.

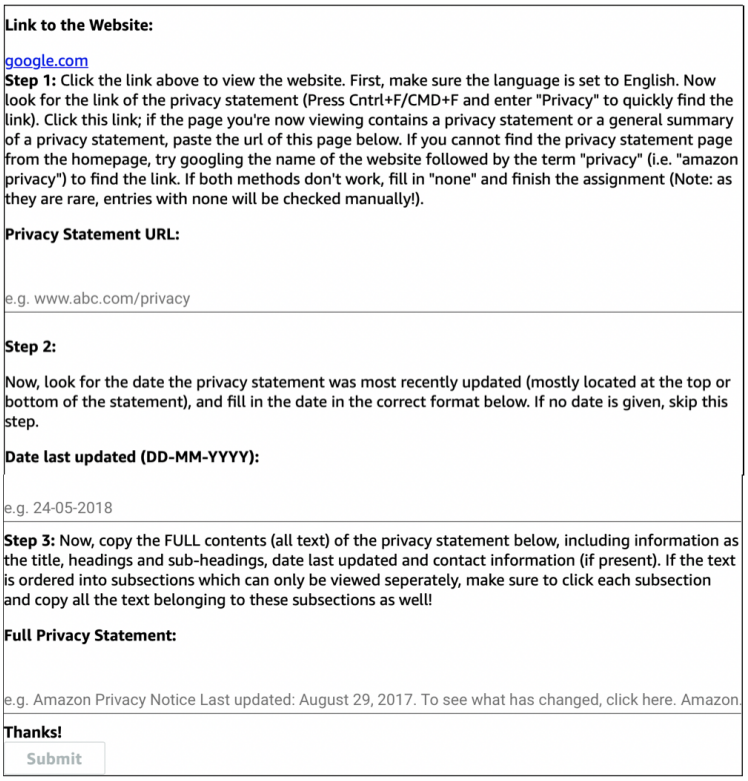


Figure 3.1: Amazon MTurk Assignment

The MTurk assignment was posted on the 10th of April, 2019. After posting the MTurk assignment, all 6000 hits were completed after 5 days and over half of the hits were completed in the first three hours (see image 3.2). This lead to a problem; as not every website contains a privacy statement, workers also had the option of filling in the option 'none' for a website. To ensure this was only done if the website did in fact

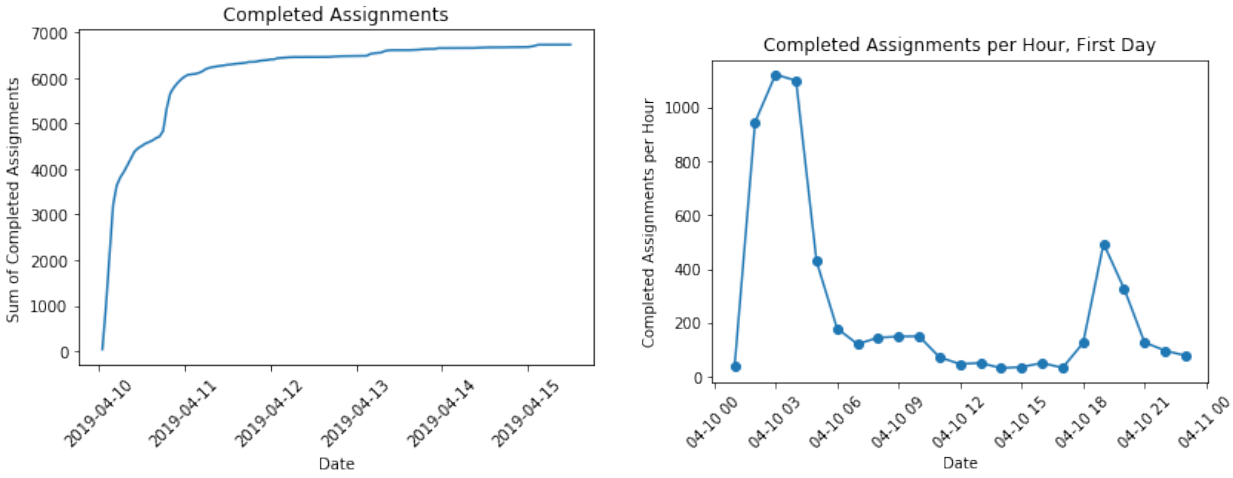


Figure 3.2: Completed Assignments

have no privacy statement, the original idea was to manually check every entry were no privacy statement was obtained. This would have to be an active procedure, as rewards were automatically approved if not manually rejected before three days. However, as over half of the assignments were completed in the first three hours and the number of 'none' entries was higher than expected (approx. 1000 of the 6700 entries), this check could not be performed as rigorous as originally anticipated. Therefore, 'none' entries were only checked if of the multiple assignments for the same website, one or two of the entries were a 'none' value. If all three entries were 'none' values, the assumption was made that the website had no privacy statement. If the three entries were not aligned, a check was performed and the incorrect entries were rejected. This lead to a total of 744 rejected assignments.

Ultimately, this resulted in a file provided by MTurk in csv-format, containing all input data.

3.3 Dataset Extraction

From the resulting csv-file, two subsets are created to use for initial testing and first results. These subsets are obtained using python, of which the full code and method can be found in Appendix F. In this section, a summary is given of how the applied code works. Image 3.4 shows a schematic overview of the process in this section.

Firstly, a number of pre-processing methods are applied. The data is checked to actually contain a string value (i.e. text input, not blanks or numbers), as a number of inputs were blank. Second, the word count of the input is set to a minimum of 9 words (some users for instance entered 'could not find link' in stead of 'none'). The shortest privacy statement encountered was 9 words (from a website called mozillazine.org), which clarifies the minimum of 9. Third, a python package called **LangDetect** is used, which analyses a text and finds which language the privacy statement is written in. The data contained privacy statements written in English, Spanish, French, Italian, German and Mandarin-Chinese. All non-English privacy statements are then removed. And lastly, the full text is checked to see if it contains the word 'privacy'. All entries without this word (often the terms of service) are removed.

From the preprocessed dataset two subsets are acquired. These two testing subsets are selected based on the assumption that if multiple workers entered the exact same input for the same website, the input is highly likely to be correct, as the chance of two workers independently making the exact same mistake is small. The small testing subset consisted of 128 privacy policies, and consisted of the websites where all three inputs

	input	url	date	full	type	wordCount	language
0	000webhost.com	https://www.000webhost.com/privacy	0001-01-01	{HOSTINGER. a Cyprus private limited company, ...	<class 'str'>	4704	en
1	000webhost.com	https://in.000webhost.com/privacy	NaN	000WEBHOST.COM - PRIVACY POLICY RESPECTS YOUR ...	<class 'str'>	4713	en
2	000webhost.com	https://www.000webhost.com/privacy	2001-01-01	. INTRODUCTION {HOSTINGER. a Cyprus private li...	<class 'str'>	4822	en

Figure 3.3: Example of three data entries for one website

(privacy statement url, last updated date and full privacy statement) where an exact character for character match. This subset is used to test the NLP-algorithms, created in section 4, and to test the code used for the analysis of the data in section 5. The second subset uses the same rule, but also adds websites where two of the three inputs were an exact match. This resulted in a subset of 701 privacy statements. This subset was used to obtain primary results, which could give an indication of promising results for the full dataset.

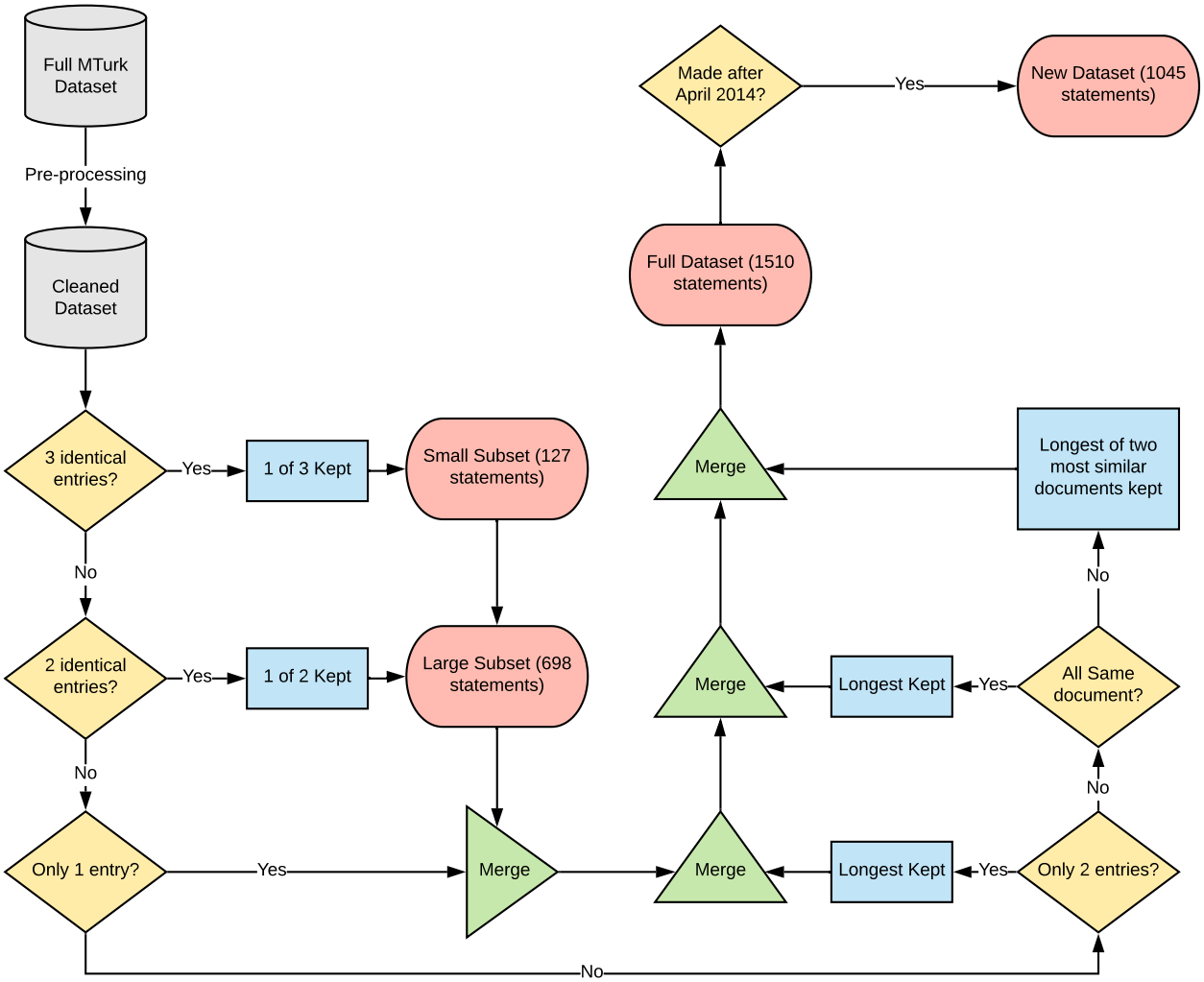


Figure 3.4: Overview of the creation of the 4 datasets

For the full dataset, the remaining websites are analyzed. These websites all had three different inputs from

the workers assigned to the websites. These differences could vary from all three copying the same document but a different part of each document (i.e. one worker could have copied the title of the statement, the next may have not copied the title but the full body of text and the last missed the first line of text) to all three workers copying a completely different document (i.e. one worker copied the privacy statement, the second copied the terms of service and the last copied the general summary of the privacy statement.

To attempt to gain the highest quality data, a number of general rules are defined and applied to all these cases. The rules were realized by manually checking a large number of inputs for each website, seeing how each input differed and what input ultimately would be most favourable to have in the dataset. For each website in the remaining dataset, there were either 3, 2 or 1 entries left after preprocessing. If only one entry was left, that entry was simply used. If two entries were left, the longest was used, as the longest often contained the full document.

If there were three entries left, a number of rules were applied. Firstly, the code assessed if the documents entered are in fact the same document, or totally different documents (i.e. the example of the privacy statement and terms of service). This is done in two steps. First, the code assesses if one or two of the inputs are significantly longer than the others. If that is the case, often one worker copied the general summary of the privacy statement, while the other copied the full text. In this case the longest entry is kept. A next step is checking how similar the inputs are. This is done by using the package **difflib**, which contains *SequenceMatcher*. SequenceMatcher is used to assess the similarity of two segments of text, and results in a score between 0 and 1. The similarity between each input is measured to assess if the documents are the same. The longest of the two documents that scored the highest similarity is then kept.

After applying the rules described above, the dataset contained 1792 entries. After dropping all duplicate entries, a total of 274 entries were removed. The amount of duplicates is high because many websites which are subsidiaries of other share the same privacy statement. After this step, manual assessment still found a small number of errors, for instance articles on websites concerning privacy, that weren't actually privacy statements. In all of the cases, the url entered by the worker did not contain the word privacy. To accurately remove these, a second check was done whereby entries of which the 'statement' didn't contain either 'privacy policy', 'privacy statement' or 'privacy policy', and the url didn't contain the privacy were removed.. This resulted in the final complete dataset, which contains 1518 of the 2000 entries, meaning 76% of the websites yielded a usable and unique result.

3.4 Post-hoc Additions

After the dataset of 1510- statements is yielded and before the Natural Language Processing (NLP) analysis is applied to the dataset, a number of post-hoc additions (or pre-NLP additions) are performed on the dataset. Firstly and most importantly, the category of each website is added to the url. This is done by using Webshrinker's[2] *Website Category API*. Webshrinker is a services company specializing in information regarding websites. It's category API is able to analyze a website and provide a corresponding category, and the certainty of it's choice for a category. Webshrinker's Website Category API is applied to all 2000 urls initially uploaded to MTurk. To assess the accuracy of the categories, a manual assessment is done of 10 random urls which had a category assigned with a certainty of less than 0,1 (for a score ranging from 0 to 1). From this manual assessment, 6 of the 10 websites provided an accurate result. As this result is acceptable for the bottom 10% and assuming this accuracy improves as the certainty improves, the categories assigned are accepted for all urls. Ultimately, the list of 2000 urls with corresponding categories is merged with the full dataset, on the input url. This enriches the data with the website category, which provides the possibility of performing the categorical analysis in section 5.2.3. A second post-hoc addition is that the top-level domain (TLD) is extracted from the website where the privacy statement is extracted from. Initially, the TLD was extracted from the input website, i.e. the homepage of the website. However, has some links of privacy

[2] www.webshrinker.com

statements redirected the user to a new website with a different TLD, the TLD is extracted as a post-hoc addition. This was done by creating a list of TLDs, and checking if a TLD from the list was present in the privacy statement url. Of all privacy statement url, a number of urls where not immediately recognized and where added later manually (for instance '.site','.dhl','.kpmg' and a few others). The last addition is that of the variable *monthsOld*. This variable calculates the age of the statement in months, from the moment the statement was acquired (April 2019). This number is calculated by translating the date input into a recognizable time format for Python, and calculating the difference in months between April 2019 and the given date. A logical drawback of this variable is that if no date was given, the age in months cannot be calculated.

Chapter 4

Natural Language Processing

4.1 Extracted variables

Table 4.1 contains all variables extracted for this project, grouped by different types of variables. For each variable, this section provides an explanation of how the variable is extracted from each privacy statement. Section 4.2 and Appendix F give further insights into how the variables are extracted. The variables chosen where based on the research review in section 2.3.1m 2.3.2 and 2.3.3. For each group of variables, an assessment was done evaluating if and how the variables could be extracted within the scope of this research. An overview of all resulting variables is given in table 4.1.

4.1.1 Variable Categories

This section gives an overview of all variables, grouped by different categories. The categories *GDPR* and *Coverage* are inspired by the research of Linden et al. (2018), the categories *Data Collection* and *Data Sharing* are inspired by the research of Harkous et al. (2018) and the vagueness metrics from the *Text Classifiers* category are inspired by Bhatia et al. (2016). The *Company Info* category concerns a number of additional variables deemed to be relevant for this research.

Text classifiers

To measure the text classifiers, each privacy statement is separated by word and by sentence, using Python's native command *split()* and the NLTK command *sent tokenize*. Once these are defined, *wordCount* and *sentenceLength* are calculated using standard Python command *len* to measure the number of sentences and the number of words, providing the number of words (wordCount) and the average sentence length (sentenceLength).

The complexity of the texts is calculated with the *fleschScore* and *fleschLevel* variables. These are based on the Flesch-Reading-Ease (FRE) score, as described in section 2.2.2. After comparing the FRE score with the Gunning-Fog index and SMOG-index (see section 2.2.2), the FRE score provided the largest

The variable *fleschScore* is calculated using the formula specified in figure 2.7. The variable *fleschLevel* is the readability level attributed to the Flesch Score. Both variables are calculated using the Python package 'Readability'. An overview of the relationship of Flesch-Level to Flesch score can be seen in figure 2.8.

Table 4.1: Table of Extracted Variables

Group	**Data Type**	**Explanation**	**NLP-technique**	**Variable Names**
Text Classifiers	Numerical, Catagorical	How and what language is used	Word- and sentence tokenization	wordCount, sentenceLength, fleschScore, fleschLevel, vagueness, conditionality, modality, generalization, numericQuantifier
GDPR	Boolean	To what extent is GDPR or are GDPR-related terms mentioned	Word tokenization	vitalInterest, legitimateInterest, publicInterest, contractualNecessity, legalObligations, unambiguousConsent, accessRights, dpo, dataController, dataProcessor, afterGDPR
Coverage	Boolean	The basic topics for privacy statements (Wilson et al., 2016)	Word tokenization	thirdParty, choices, security, specificAudiences, dataRetention, policyChange
Third Parties	Boolean	Check if and what third parties are mentioned	Word- and Sentence Tokenization, Named-entity recognition	privacyShield, nai, daa, google, facebook, amazon, microsoft, criteo, adobe, doubleClick, nielsen, paypal, apple, stripe, yahoo, android, socialMedia
Data Collection	Boolean	What types of data are collected	Word- and sentence tokenization, dependency parsing	computerInformation, contactInformation, webTracking, personalInformation, financialInformation, location, demographicInformation
Data Sharing	Boolean	What types of data are shared	Word- and sentence tokenization, dependency parsing	generalShare, dataCombining, computerInformationShare, contactInformationShare, webTrackingShare, personalInformationShare, financialInformationShare, locationShare, demographicInformationShare
Company Info	Catagorical, Boolean, Numerical	Other company-related information	Word tokenization, Named-entity recognition	addressProvided, emailProvided, headquarters, monthsOld, TLD, cat, changeNotification

Lastly, the variables *vagueness, conditionality, modality, generalization* and *numericQuantifier* are based on 'A Theory of Vagueness and Privacy Risk Perception' (Bhatia et al., 2016), which is reviewd in section 2.3.1. This research focusses on defining words which cause vagueness within privacy statements. The vague

words found in this research are divided into four categories, namely conditionality words, modality words, generalization words and numeric quantifiers. Together, these groups of words lead to vagueness in privacy statements. To calculate these variables in this research, the *Term Frequency* method explained in section 2.2.1 is used. Lastly, the numbers are multiplied by 1000 in order to provide a more readable number.

GDPR

The GDPR variables are variables which are triggered when certain words are present in the text. The words are all related to GDPR and function to give an indication to which extent a company attempts to take GDPR into account. For instance, if a privacy statement is updated around or after the entry-date of GDPR (25 May 2018), and the statement discusses data controllers, -processors, legal bases for gathering data en the GDPR itself, one can assume that this company has adjusted their privacy statement for GDPR. If none of these are true, the opposite is more likely. The first six variables (vitalInterest to unambiguousConsent) concert the six bases for processing personal data from the GDPR. The GDPR states in article 13 that upon collecting personal data, the data controller (i.e. the company) shall provide: *the purposes of the processing for which the personal data are intended as well as the legal basis for the processing* (European Commission, 2017). As many statements collect some form of personal data, these variables can give an indication to what extent the legal bases are mentioned. Variables *accessRights* to *afterGDPR* concern if specific terms are mentioned which are directly related to the GDPR. Lastly, *afterGDPR* concerns if the privacy statement was update around or after GDPR came into effect, which is true if the statement is updated after February 2018.

Coverage

Coverage concerns if the basic topics of privacy statements are mentioned, which are defined in 'The Creation and Analysis of a Website Privacy Policy Corpus' (Wilson et al., 2016). Similarly to the GDPR variables, these variables are triggered if a certain word or combination of words is found in the statement. For each variable an their corresponding trigger words, see section 4.2, and for more elaboration on the working of the Python code, see appendix F. The paper by Wilson et al. (2016) also contains 3 categories which are not covered in this research, namely 'First Party Collection/Use', 'Do Not Track' and 'Other'. The first category is already covered in the 'Data Collection' group in this research, specified below. The 'Do Not Track' (DNT) variable concerns a browser configuration which has been proposed, which would in theory automatically block all user web-tracking from third parties (Rosch, 2011). However, due to insufficient deployment and support, further advancement of this project has been cancelled (W3C, 2019). Furthermore, upon manual validation it was found that although many statements mention the DNT signal, the most frequent response to the signal was that the website did not work with the DNT-option. This report therefore chooses to not include this variable. Lastly, 'other' concerned an additional, undefined group of miscellaneous topics. As the variable is unspecified, it cannot be used in this research.

Third Parties

The 'Third parties' group contain binary frequency variables (see section 2.2.1) which are 'True' when the privacy statement mentions the third party. For instance, if Facebook is mentioned anywhere in the privacy statement, the variable *facebook* is set to true. The selection of third parties was done by using *Named-entity recognition* (NER) from the Spacy Python package. NER is a method to classify named entities into pre-defined categories, one of which being organizations. For the small subset, a list of most frequent occurring organizations was created. From this list, the most frequently occurring third parties were extracted, with a cut-off point at five or less occurrences. Similarly to the previous groups, trigger words are then defined for each variable. An important implication here is that different third-parties can have different implications for

a privacy statements attitude towards privacy. For instance, privacy statements mentioning *Privacy Shield*, or the *Network Advertising Initiative* could have different implications than privacy statements which mention Facebook or Google.

Data Collection

Data collection concerns a number of binary variables, which are each set to 'True' if the specific type of data is collected. The types of information are taken from Harkous et al. (2018). To assess these variables, each sentence is individually checked if the topic of the sentence concerns collectind data. This is done by using with dependency parsing (introduced in 2.2.1. To fully explain how this works, consider the following example, taken from the privacy statement of Tribune Publishing.

Example 4.1.1. ***We may collect and analyze information such as IP addresses.***

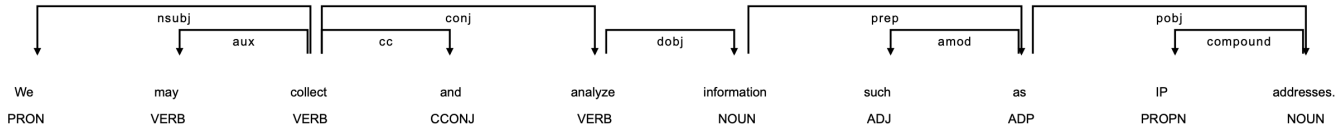


Figure 4.1: Dependency parsed sentence from example 4.1.1. This example triggers the algorithm and sets the variable "Financial Information" to "True".

For this sentence, the algorithm first checks if one or more pre-defined trigger-words are present in the sentence. In example 4.1.1, the combination of 'IP adddresses' is a trigger-word for the variable *computerInformation*. Next, the words of the sentence is lemmatized (as defined in section 2.2.1), and the sentence is searched for verbs which are listed as trigger-words attributed with data collection (specified in section 4.2. As can be seen in figure 4.1.1, the words 'collect' and 'analyze' are recognized as verbs. Lastly, dependency parsing is used to ensure that the data collection trigger-word does not have any negative dependencies. In example 4.1.1, this is not the case. Thus, the variable *computerInformation* is True, as the privacy statement explains that the company collects computer information.

To illustrate the case of negative dependencies, an additional example is provided, taken from the privacy statement of *www.livestream.com*.

Example 4.1.2. ***We do not directly collect or store credit or debit card numbers.***

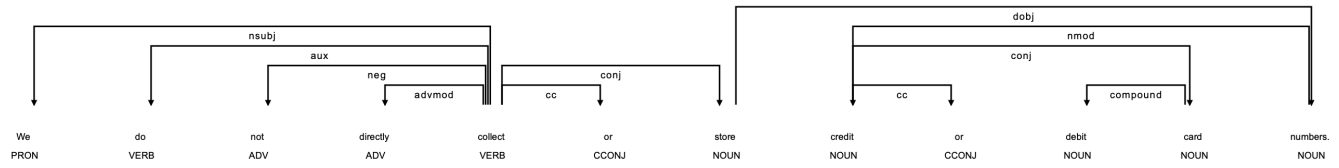


Figure 4.2: Dependency parsed sentence from example 4.1.2. This example does not trigger the algorithm, and keeps the variable "Financial Information" "False".

In the example above, the sentence contains words which are trigger-words for the variable *financialInformation* (namely 'debit card'), which concerns if the privacy statements explains the collection of financial information. Also, the sentence contains a verb relation to data collection, namely the verb 'collect'. However, due to the negative relation of the verb 'collect' with the word 'not', the variable fiancialInformation is kept 'False'.

Data Sharing

The algorithm for extracting what types of data is collected works identically to the algorithm extracting which types of data are collected, only now verbs which are synonyms of sharing trigger the variables. For an overview of these words, see section 4.2. The group contains two variables which do not concern what types of information is shared, namely *generalShare* and *dataCombining*. The first variable assesses if the privacy policy explicitly states that information is shared, without specifying what kind of information. As every website automatically collects certain information upon visiting, no such variable is added for the 'Data Collection' group. The second variable, *dataCombining*, concerns if the company combines their data with third-party data sources. See for instance example 4.1.3.

Example 4.1.3. ***We or our third-party partners may combine information that each of us collects automatically with other information about you, including information you choose to provide.***

To assess if the sentence regards combining data with third parties, a similar method is used to the method described above on the variables concerned with data collecting and -sharing, namely dependency parsing. First, it is assessed whether the lemmatized words in a sentence contain the word 'combine' or a synonym of this word. Subsequently, the sentence is checked for words which can indicate that the subject of the sentence is third parties. Lastly, the verbs of the sentence are checked to see if the verb 'share','collect' or any synonym of these words is used in the sentence, without negative dependencies. Looking at example 4.1.3, the word 'combine' is the first verb in the sentence. The words 'third-party' and 'partners' are both trigger-words that the sentence regards third parties. Lastly, the sentence contains the verb 'collects', without negative dependencies. Combined, the algorithm now concludes that this sentence is about combining data with third parties.

Company Info

The Company Info group concerns variables directly related to the company and the privacy statement. The TLD variable concerns the top-level domain of the website (e.g. .com, .net, .org), and is extracted from the input url of the website. The *monthsOld* variable concerns the age of the privacy in months, calculated from March 2019 (the month of extraction of the privacy statements). The variables *addressProvided* and *headquarters* concern a binary variable if an address can be found within the privacy statement, and secondly where the headquarters of the organization are located.

Both variables use methods which do not ensure full accuracy, but do score enough accuracy to yield usable results. For both variables, the package **country list** is used to generate a list of all countries in English. From this list, for each privacy statement the freuqency of occurrences of each country is counted. The country with the most occurrences is listed as the headquarters. For the address, each sentence is checked if it contains a country and a number. If it does, this sentence is recognized to contain an address, and the variable is set to true. If the statement contains no such sentence, an additional check is done using the package **pyap**, or *Python address parser*. This package recognizes U.S. and Canadian addresses in a string. This method is used as additional check, and not as primary check, to ensure faster run time. If this method also yields no result, the *addressProvided* variable is set to 'False'.

4.2 Validation

This section discusses the manual validation of each variable extracted. Due to the limitations and scope of this project, a relatively simple form of validation is chosen to assess the validity of each metric. After each

method to extract the variable is chosen, a Python script is created to extract ten random privacy statements and their scores on the variable (the number ten was chosen here as it is seen as an adequate amount for this research, and an easy number to apply a score to). Based on manual validation of these ten privacy statements and their respective scores, a quick indication can be given on what the performance is of each variable. If the variable scored lower than 70% accuracy, a reiteration was done where rules were added to the Python code, and the validation was repeated on ten new random privacy statements. In this section, the last iteration of each round is discussed. Per group of variables, a short review is given on each variable and their performance. This type of manual validation naturally has limitations; these will be discussed in the discussion, chapter 6.

Note that the manual validity test will not be done for all variables. For instance, variables from the group of 'Text Classifiers' are not validated due to the nature of these variables. The variable *wordCount* can always be deemed to be accurate as it requires Python to perform the simple task of counting all words. Also, for the variable *vagueness*, no real validation can be done as vagueness is a subjective concept (Bhatia et al., 2016). All variables not validated are stated at the end of this section, with the corresponding reasons why.

This section discusses 'trigger-words'. Trigger-words are one or multiple words which trigger a variable. This is assessed by converting the sentence to all lower-case letters, and searching if the text contains the same combination of letters as the input (similar to using the search-function on a website or document). The symbols used in the trigger-words column in the tables are described in table 4.2 below.

Table 4.2: Variable Validation Legend

Symbol	**Meaning**
'	Beginning or end of search-term
[]	All words between brackets need to be present in one sentence.
/	Or (one of the values before and after symbol needs to be met).
&	And (both requirement before and after symbol need tobe met)
*italic font with asterisk**	Requires additional information, provided in footnote below table.

Table 4.3: GDPR-Variable Validation

Variable	**Accuracy**	**Trigger-words**
legitimateInterest	100%	'legitimate interest'
contractualNecessity	90%	'contract'
vitalInterest	100%	'vital interest'
publicInterest	90%	'public interest'/[*Share Word**, 'authority']
legalObligation	100%	'legal obligations'
unambiguousConsent	80%	[['legal','bases']/ ['legal','basis']/ ['lawful','bases']/ ['lawful','basis'] & 'consent']
accessRights	100%	['right','access']/ ['right','request']/ ['right','information']/ ['request','access']

* *Share words are a selection of lemmatized words which are synonyms of the word 'share'. To assess this, each word is lemmatized and checked with the list of share words. The words are lemmatized in order to ensure each form of the verb can easily be identified.*

GDPR

This group of GDPR-related variables scores relatively high accuracies, as they extract information related to specific terms. For instance, extracted if a company explicitly mentions legitimate interest in their privacy statement is relatively easy, as it is an unambiguous term. Even the word 'contract' indicated in 9 out of 10 validations a contractual necessity for processing personal information. For unambiguous consent, however,

the word 'consent' alone only yielded an accuracy of 40%. Combining this with a combination of words regarding the lawful bases of processing increased this accuracy to 80% (Although this technique could be used for every variable, the incorrect 20% consisted of false positives, as not every sentence necessarily explicitly states it concerns 'lawful bases' or 'legal bases' when explaining unambiguous consent. Therefore, if possible, this method is avoided). For public interest, the variable first yielded an accuracy of 60% when only searching for public interest. Adding the combination of a synonym of the word 'share' and 'authority' (or authorities) increased this accuracy to 90%. This was done as many companies indicated that personal information would be shared with authorities if deemed necessary, which essentially is the same as public interest.

Table 4.4: Coverage-Variable Validation

Variable	**Accuracy**	**Trigger-words**
thirdParty	100%	'third party'/'third parties'/ 'third-party'/ 'third-parties'/ 'business partners'
choices	100%	['opt in','opt out','opt-in','opt-out']/ ['your','choice'], ['you','choice'], ['you','choose']
security	90%	'security'
specificAudiences	100%	'children'/'under 13', 'california privacy rights'/ 'coppa'
dataRetention	90%	['data','retention']/ ['information','retention']/ ['long','keep','data']/ ['long','keep','information']/ ['retain','data']/ ['retain','information']
policyChange	100%	'policy'/'statement'/'notice' & 'change'/'changes'/'update'
changeNotification	90%	'notice'/'notify' & 'e-mail address'/ 'e-mail'/ 'email'/ 'e mail' & *additional rule*[*]

[*] *The additional rule first checks if the policy contains a sentence regarding policy change. Following this sentence, the following seven sentences are analyzed to assess if they contain words similar to 'notify' and 'e-mail'. The number seven was chosen after multiple iterations of the script to test which number of sentences yielded the highest accuracy without leading to false positives (for instance, the option to notify the company via e-mail for any further questions was recognized as a false positive).*

Coverage

The coverage variables also yield relatively high accuracy with a small number of words. This could potentially also be attributed to the relatively low diversity of documents analyzed (namely all privacy statements), leading to low word-ambiguity between the documents. A small number of occasions reduce the accuracy of the variables. For instance for the 'security' variable, the website of www.steam.com states the following:

Example 4.2.1. ***If you pay by credit card, you need to provide typical credit card information (name, address, credit card number, expiration date and security code) to Valve...***

This triggers the security variable, while not actually discussing data security, leading to a false positive.

The privacy statement of torstar.com shortly mentions data retention, but does not use the trigger-words defined (see 4.2.2). This sentences therefore produces a false negative.

Example 4.2.2. ***Once we no longer require your personal information for purposes you have consented to, we securely dispose of it.***

Lastly, the 'changeNotification' variable also produces is a false negative in the following sentence from the privacy statement of *shopify.com*, which states the following:

Example 4.2.3. ***If we make material changes to this Privacy Policy, we will give you notice of such changes by posting the revised policy on this Website, and where appropriate, by other means.***

Although the statement does not explicitly specify what 'other means' entails, one can assume that some form of direct communication is implied. However, this false positive can also be attributed to the ambiguity of the text.

Data Collection and Data Sharing

For the next two tables, concerning the 'Data collection' and 'Data Sharing' group, the NLP-techniques are more complex than simply assessing if certain trigger-words are present. Although these variables also use trigger-words, more requirements need to be met before the variable is set to positive. For elaborate explanation of the working of this code, see Appendix F.

Table 4.5: Data Collection Variable Validation

Variable	Accuracy	Trigger-words
location	100%	'location'
computerInformation	100%	'operating system','browser type','system type','unique device','browser software', 'ip address', 'internet protocol','device information', 'ip-address'
demographicInformation	70%	' gender ',' age ','education','profession','occupation','income level','marital status'
contactInformation	100%	'contact information', ' name', 'email address', 'home address', 'postal address', 'country of residence', 'first name', 'last name', 'e-mail address', 'telephone number'
webTracking	100%	' logs ', 'beacons', 'tracking cookies', 'device fingerprinting', 'tracking methods', ' trackers '
financialInformation	90%	'credit card', 'billing information', 'debit card'
personalInformation	100%	'personal information', 'personally identifiable information', 'pii', 'personal data'
dataCombining	80%	'combine' & 'third party','third parties','third-party','third-parties'

From this section on, more complex methods are used which yields a slightly more varying accuracy per variable. Furthermore, for these cases, all variables use largely similar methods to extract information. Because of this a distinction can be made between errors found which are created due to the method used, i.e. errors which are possible for all variable, and errors which are specific to that variable.

First the errors regarding the method used are highlighted, i.e. errors which can potentially occur for each variable. For the variable 'demographic information', the privacy policy of sfgate.com provided en error. This was due to the structure of their privacy statement; different types of data collection were listed, whereby each bulletpoint ended with a full-stop. For example:

Example 4.2.4. ***Information we collect...may include details such as: ... •Demographic, interests and household information (e.g., age, gender or education).***

As this bullet was the second in the list and the first bullet also ended in a full-stop, this section (Demographic,.. education.) was recognized as a separate sentence. As this sentence does not contain any word resembling collection, it was not recognized and provided a false negative.

A similar error was also recognized for the data combining variable in the privacy statement of *costco.com*. The privacy statement uses two sentence to explain what kind of data it receives from third parties, and subsequently explain that that information is combined with their own information. As the algorithm checks sentence per sentence, this resulted in a false negative.

Another method related error can be attributed to the recognition of a negative word in the sentence, which in reality is not related to the collection of data. For instance:

Example 4.2.5. ***Our Applications are flexible and allow our Customers to collect a variety of personal data from and about their Customer Business Contacts, including name, organization, title, postal address, e-mail address, telephone number, fax number, social media account ID, credit or debit card number and other information including but not limited to dietary preferences, interests, opinions, activities, age, gender, education and occupation.***

This example states the collection of a large number of different types of information. The second part of the sentence contains the word 'not' however, which is recognized as a negative word. Because of this, the algorithm skips this sentence. The possibility of this error was taken into account while writing the code; although the code first used dependency parsing to analyze which negative word was grammatically connected to the verb which described the collection, manual validation showed that often the dependency parser failed to accurately recognize which negative words belonged to which verb (for instance, the sentence 'We do not collect demographic data' would incorrectly set the demographic variable to true, because the algorithm did not recognize the relation between 'not' and 'collect'. This example is fairly simple, in reality this mostly occurred with more complex sentences). This created many false positives. Changing the algorithm to simply look for a negative word in the sentence resulted in higher accuracy, as the previously described misclassification was more prevalent than example 4.2.5 listed above.

There were also some errors to due vagueness of the text. For instance, while the privacy statement of *oath.com* does not explicitly state collecting demographic data, it does state sharing this data. As data which is shared must also be collected, one can deduce that this data is collected but the NLP-algorithms cannot. Another example of this vagueness could be found in the privacy statement of *domaintools.com*, for the variable of data combining. While the statement does not explicitly state combining third-party data with their own data, it does state collecting personal data directly, and also receiving personal data from third parties. From this it can be assumed that this data is combined, although this is not explicitly stated.

In comparison to extracting what types of data are collected, extracting the types of data which are shared is more complex. Privacy statements less frequently explictly state what types of data are shared, but limit their explanations to why the data is shared and with what 'category of recipient', as stated in the GDPR (European Commission, 2017). The GDPR does however not obligate a company to share exactly what types of information are shared. Because of this, many companies choose not to disclose this information, and when it is disclosed there are no standard practices which are applied. Therefore, to ensure valuable information is retrieved from the statements, in some cases a extra step was also taken to measure the true-positive rate (TPR). The TPR consists of the percentage of positive values which is accurate. For instance, upon measuring the accuracy of *locationShare*, the three errors found were all positives. This lead to checking all 22 positive values found in the subset of 150 statements, were only 10 were accurate. Thus this variable must be used with caution, as for instance grouping companies based on this variable might lead to inaccurate conclusions. However, as overall accuracy is still 70%, this variable is taken into account for the research. For the *demographicInformationShare* variable, only 4 positives were found. The validation test set consisted of 10 'False' values. Therefore, the positives were also manually assessed which yielded a TPR of 75%, which is deemed to be sufficient.

As with data-collection, the method of retrieval also caused some errors, i.e. some errors were found which are possible for each variable. For instance this statement from the privacy statement of *twitter.com*;

Table 4.6: Data sharing variable Validation

Variable	Accuracy	Trigger-words
generalShare	100%	['information','data'] & ['share','disclose']
locationShare	70% , TPR:45%	'location'
computerInformationShare	100%	'operating system','browser type','system type','unique device','browser software', 'ip address', 'internet protocol','device information', 'ip-address'
demographicInformationShare	100%, TPR:75%	' gender ',' age ', 'education', 'profession', 'occupation', 'income level', 'marital status'
contactInformationShare	80%	'contact information', ' name', 'email address', 'home address', 'postal address', 'country of residence', 'first name', 'last name', 'e-mail address', 'telephone number'
webTrackingShare	100%	' logs ', 'beacons', 'tracking cookies', 'device fingerprinting', 'tracking methods', ' trackers '
financialInformationShare	80%	'credit card', 'billing information', 'debit card'
personalInformationShare	100%	'personal information', 'personally identifiable information', 'pii', 'personal data'

Example 4.2.6. ***You can choose to share additional information with us like your email address, phone number, address book contacts, and a public profile.***

As the sentence contains sharing, certain types of data and no negative words, the algorithm concludes that these types of data are shared. However the sentence explains that the user shares their data with Twitter, essentially explaining data collection as defined by this research. As few companies choose to use this formulation, this is accepted as a limitation of the algorithm.

In other situations companies to not explicitly mention sharing types of data, but from the third parties the type of data can be derived. For instance this example from the privacy statement of *netflix.com*:

Example 4.2.7. ***For example, we engage Service Providers to provide marketing, advertising, communications, infrastructure and IT services, to personalize and optimize our service, to process credit card transactions or other payment methods, to provide customer service, to collect debts, to analyze and enhance data (including data about users' interactions with our service), and to process and administer consumer surveys.***

Here, Netflix explains with whom personally identifiable information is shared. As these 'Service Providers' are engaged 'to process credit card transactions', it can be concluded that financial information is shared. However this is not explicitly mentioned, providing a false negative.

Table 4.7: Company Information Variable Validation

Variable	Accuracy	Trigger-words
addressProvided	100%	none (see appendix F for full code)
emailProvided	90%	none (see appendix F for full code)
headquarters	100%	none (see appendix F for full code)

Company Info

Lastly, for the company info variables three variables can be assessed for accuracy. No trigger-words are

given as these algorithms do not work with trigger-words. The variable *addressProvided* works by a combination of methods, namely an external package which recognizes some addresses, combined with a package which recognizes countries. The variable *emailProvided* works by searching for a word which contains an at-symbol (@) and a full-stop within the word. Lastly, *headquarters* works by identifying all countries mentioned in a privacy statement, and selecting the most frequently mentioned country. Although this method would seem prone to errors, after testing multiple methods (i.e. also looking for addresses, or looking for (synonyms of) the word headquarters) this method yielded the highest accuracy. As can be seen in table 4.7, these algorithms are quite accurate in extracting correct information. The one error found for the variable *emailProvided* was due to an e-mail address being hyperlinked in the privacy statement, and not explicitly written down.

In the last table, the variables are named which cannot be tested for accuracy and why.

Table 4.8: Variables with no validation

Variable	**Reason**
wordCount, sentenceLength	These are calculated by Python and the NLTK package.
fleschScore, fleschLevel	Calculated using a formula (Flesch, 1949)
vagueness, conditionality, modality, generalizaton, numericQuantifier	Indicates the presence of a selection of words, calculated using selection of vagueness-indicators (Bhatia et al., 2016)
privacyShield, nai, daa, google, facebook, amazon, microsoft, criteo, adobe, doubleClick, nielsen, paypal, apple, stripe, yahoo, android, socialMedia	Assesses if the name of the company or organisation is mentioned in a privacy statement.
monthsOld	Calculated based on user input. Accuracy depends on quality of the entries, not quality of the code.
TLD	Extracted directly from the privacy statement url.

Possible accuracy dips here are due to external errors from external packages (i.e. NLTK incorrectly parsing a text into seperate words, causing wordCount to vary). However, inaccuracies are deemed to be small and unlikely, and to have no significant impact on results.

4.3 Conclusions

Based on the applied method of variable extraction and validation, some conclusions can already be given. This section provides an overview of these conclusions, and other findings based on the research in this chapter.

First of all, a general statement that can be made is that even though relatively simple NLP-methods are used, high accuracy's are achieved. A benefit of only analyzing privacy statements is that the text is less ambiguous than when for instance analyzing a whole range of different documents. Therefore, when the word security is used in a document in this research, an assumption can be made with high accuracy that the sentence regards data security. When analyzing a large variety of documents, this is not the case, as security can be used in a variety of different ways. What can be concluded from this is that because the documents being analyzed in this research are all similar (i.e. are all privacy statements), simple methods of NLP can be used.

A second conclusion that can be made is based on the terms which relate to variables. When measuring the accuracy of the variables, the less ambiguous the terms related to the variable are, the higher the eventual

accuracy. This can for instance be concluded from the GDPR legal-bases variables. 'Legitimate interest' is an unambiguous term, and the only term used to indicate legitimate interest. Because of this, finding out if a company uses legitimate interest as a legal basis is simple, which can be seen from the high accuracy which is yielded in the validation. This is not only true for this research, but also in reality; if a user of a service is looking for a specific term in a privacy statement which can be only noted in a single way, the chance of finding the term and therefore informing the user increases. However, if there are multiple ways to reference the term, the user will find it more difficult to find what he or she is looking for.

Chapter 5

Data Analysis

This section discusses the analysis of the dataset extracted from the privacy statements. Which variables are extracted, how they are extracted and what their meaning is is discussed in section 4. A full report on the values of each variable is given in appendix E.

The section will start with an overview of the data and it's attributes in section 5.1. Following this, a number of analyses will be performed in section 5.2. The presence of clusters is assessed in section 5.2.6. Lastly, based on the findings in the previous sections, conclusions and implications are given in section 5.4.

Before the analyses were performed, based on the results from the NLP tests a few entries were dropped. Even though in section 3 the longest entry was chosen, the data contained a small number of summaries of statements, in stead of full statements. Also, some entries did not only consist of the privacy statement, but also the full terms of service of the companies. Lastly, for two entries, the NLP packages were not able to recognize two large tables, leading to the algorithms thinking the tables were one long sentence. This resulted in two Flesch-scores of approx. -50. These anomalies were removed, which resulted in dropping 8 policies, leading to a final dataset of 1510 statements.

5.1 Data Overview

As discussed in section 3, the full privacy statement dataset consists of 1518 privacy statements, which combined with 71 variables provides a dataset of 107210 values. To obtain a more tangible understanding of the dataset, this section provides an overview of the characteristics of the variables yielded from the privacy statements.

5.1.1 Old- and New Dataset

The creation of the *monthsOld* variable in section 3.4 makes it possible to divide the dataset based on the age of the privacy statement. Doing this makes it possible to analyze for the effect of age of statements, i.e. if it matters if a company updates it's statement or not. Most importantly however, it makes it possible to perform analyses with a dataset which only contains privacy statements which are updated recently, giving a better indication of the current state of privacy statements (privacy statements which haven't been updated for quite some time could be linked to a dead website, which therefore does not represent the current state of privacy statements). The division of old- and new statements is set at 60 months, or 5 years, meaning every statement updated before April 2014 is regarded as old, and every statement updated during or after April 2014 is regarded as new. The period of 5 years is chosen because of manual assessment, where frequently used

websites sometimes had valid privacy statements of more than four years old. Websites which had statements of over 5 years old often seemed outdated, or the privacy statements themselves seemed outdated. As a last remark, statements where no date was filled in are automatically added to the 'old' group, as their creation date is unsure; this is useful as the goal of the separation is to create a group of recent privacy statements, not necessarily define a group of older privacy statements.

Removing the old statements results in a database of 1045 'new' statements which are updated after April 2014 (based on the input provided by the Amazon Mturk workers). The 'old' group contains 465 privacy statements, of which 58 actually have an 'last-updated' date before April 2014; the other 407, 26.8% of the entries, all have no 'last-updated' date and therefore no conclusion can be made on the date. The size of this group naturally has implications for the method and the results of the analysis, which will be further discussed in chapter 6.

Removing the statements which do not have 'last-updated' dates makes it possible to do a short analysis on the differences between websites with a new and websites with an old privacy statement. Based on these statements with a date, 5.2% of the statements are older than 5 years, meaning the far majority of the websites has updated their privacy statement within the last 5 years. Interestingly, when comparing the average majestic rank of the old- and the new group, no significant differences can be found. This indicates that more frequently websites aren't better in updating their privacy statements. There are however interesting significant differences in region. The group of old privacy statements have significantly less European statements than the new group, a possible indication of the effects of the GDPR. Upon manually checking a number of the old privacy statements, many websites seemed dead or outdated.

For the coming analyses, each analysis will be performed for the full dataset and the new dataset. If differences between the two datasets can be identified, these will be discussed and then only the new statements will be analyzed to ensure the most recent statements are taken into account for the analysis. If no significant differences are identified the full dataset is used, and nothing additional is mentioned. This is done to ensure either the maximum amount of data is used (all privacy statements), or when necessary, only accurate data is used (recent datasets of which the data is more sure to be correct).

5.1.2 Variable Groups

Text Classifiers

The privacy statements range anywhere between 111 words (www.un.org) to 23937 words (www.turner.com), meaning that taking an average reading speed of 250 words per minute, reading the full privacy statements takes anywhere between 27 seconds to 95 minutes. Table 5.1 shows an overview of the Flesch reading analysis results. Most statements are either difficult or very confusing to read, equivalent to only being understood by college- or university graduates respectively (Flesch, 1949). Only 6 of the 1510 statements are regarded as being written in plain English.

Vagueness concerns a relative measure which indicates the frequency of vague words used in a text. The most vague privacy statement was that of *chicagosuntimes.com*, and the least vague privacy statement analyzed was that of *bangkoktimes.com*. As this variable is a self-constructed measure, it is important to check this variables for meaningful correlations, which will be done further in this chapter. Otherwise no meaning can be attributed to the values scores, other than relative scores.

GDPR

The GDPR-variables are divided into GDPR-terms which are related to the GDPR, and the legal bases, which cover the six legal bases for processing data. Of all statements analyzed, most statements (63%) where updated after GDPR. The only GDPR-topic that was covered in the majority of the statements is Access

Table 5.1: Reading Ease, based on Flesch Level

Category	**Frequency**
Difficult	61.7%
Very Confusing	35.8%
Fairly Difficult	2.0%
Standard	0.4%

Table 5.2: Coverage of GDPR-themes

GDPR-topic	**Frequency**
Access Rights	77.1%
Statement written during/after GDPR	58.7%
Data Controller	25.6%
Data Protection Officer	23.2%
Data Processor	9.5%

Rights, or the right for the user to access their data. For the rest of the variables, all eventual measured frequencies where below %50. 2.5% of the statements do not mention any GDPR-related terms at all. For the legal bases, the mos -mentioned basis for processing personal data is legitimate interest, which is mentioned in 42% of the statements. 37% of the statements do not mention any legal basis for processing personal data. See table 5.2 and 5.3 for an overview.

Coverage

The coverage variables, which assess more general topics for privacy statements which are not related to GDPR, generally yield higher scores than the variables in the GDPR group (i.e. the topics these variables represent are mentioned more often than the GDPR-related topics). For every variable the majority of the statements do tend to cover these topics, i.e. all frequencies where above 50%. This further confirms Wilson et al.'s (2016) research, that these topics are very prevalent in privacy statements.

Third Parties

Google is the most mentioned third party in all privacy statements, with a presence of 54%. This means that for more than half of the companies involved in this research, some form of transaction or cooperation is linked to Google. Second is Facebook, which is mentioned in 43% of the statements. A far third is *Android*, concerning Google's mobile operating system, which is mentioned in 12% of the statements. Furthermore,

Table 5.3: Coverage of Legal Bases for Data Processing

Legal basis	**Frequency**
Legitimate Interest	41.5%
Contractual necessity	36.6%
Legal obligation	30.4%
Public interest	27.4%
Unambiguous Consent	17,3%
Vital interest	6.1%

Table 5.4: Coverage of main topics

Topic	**Frequency**
Third parties	94.7%
Sharing information	89.3%
Data Security	89.0%
Changes of the Statement	87.8%
Your choices	78.3%
Retention of data	69.9%
Specific audiences	58.5%

27% mention the *Network Advertising Initiative* and 19% mention the *Digital Advertising Alliance*. These are organizations concerned with the fair use of data for advertising, and therefore give an indication if the company upholds to certain privacy standards. Other third parties are more likely to be mentioned because of the use of specific services for which data exchange must take place, and are therefore regarded separately. The full overview of results can be seen in table 5.5.

Table 5.5: Mentioned Third Parties

Third Party	**Frequency**
Google	54.3%
Facebook	42.3%
NAI	27.2%
DAA	19.2%
Android	12.8%
Adobe	12.7%
Apple	11.9%
doubleClick	10.8%
PayPal	8.3%
Microsoft	8.1%
Amazon	8.0%
Nielsen	5.0%
Stripe	4.9%
Yahoo	3.6%
Criteo	1.8%

Data Collection

The data collection variables yield more mixed results. Nearly all privacy statements (92%) explain that personal information is collected. The large majority of statements explain collecting contact information and computer information (86% and 79% respectively). Geo-locational data is collected in 54% of the statements, and web-tracking information is collected in 47% of the statements. The least frequent type of information collected is financial information, which is collected in 38% of the statements. Lastly,12% of the statements explicitly admit to combining their collected information with third parties. An overview of this explanation can be seen in table 5.6.

Table 5.6: Data collection over all statements

Information Type	Explanation	Frequency
Personal Information	Umbrella term, information which can be linked to an individual	91.7%.
Contact Information	E.g. name, address, phone number	86.1%
Computer Information	E.g. operating system, browser type, ip-address	79.4%
Demographic Information	E.g. gender, age, profession	66.4%
Geo-locational information	Whereabouts, unspecified granularity	52.1%
Web Tracking Information	E.g. tracking cookies, web beacons, device fingerprinting	45.2%
Financial Information	E.g. credit card number, debit card number, general billing information	37.0%
Data Combining	The statement states that data from third parties is combined with data collected by the website	11.9%

Table 5.7: Data sharing over all statements

Information Type	Explanation	Frequency
General Share	If any kind of information is shared to third parties	89.3%.
Personal Information	Umbrella term, information which can be linked to an individual	75.3%.
Contact Information	E.g. name, address, phone number	36.1%
Geo-locational information	Whereabouts, unspecified granularity	11.9%
Financial Information	E.g. credit card number, debit card number, general billing information	8.4%
Computer Information	E.g. operating system, browser type, ip-address	8.4%
Web Tracking Information	E.g. tracking cookies, web beacons, device fingerprinting	4.1%

Data Sharing

The data-sharing variables yield largely varying frequencies, which provides an insight into the informativeness of statements. 90% of the statements explicitly state sharing information with third parties, and 77% state sharing personal information with third parties. However, the exact types of information (e.g. contact information, computer information, demographic information etc.) score far lower positive-rates. In reality this is impossible; if a company admits to sharing data, some type of data must be shared. A plausible conclusion to this is that companies don't explicitly mention what types of data they share. This is also plausible when looking at the GDPR; in article 14 states that if personal information is shared, it must be notified to the user, together with the purpose of sharing and to which group (European Commission, 2017). However, the type of data which is shared does not necessarily need to be specified.

Company Information

The company info variables concern a number of variables related to the privacy statement and the company. The oldest privacy statement found was that of the Indian website of the clothing brand *Puma*, which dated from 1st of January 2004. The average privacy statement was last updated 16 months before retrieval of the dataset, with the far majority being updated in the last 20 months. Below we show two histograms, the first being of the full dataset, and the second being of the new dataset (i.e. no statements older than 5 years)

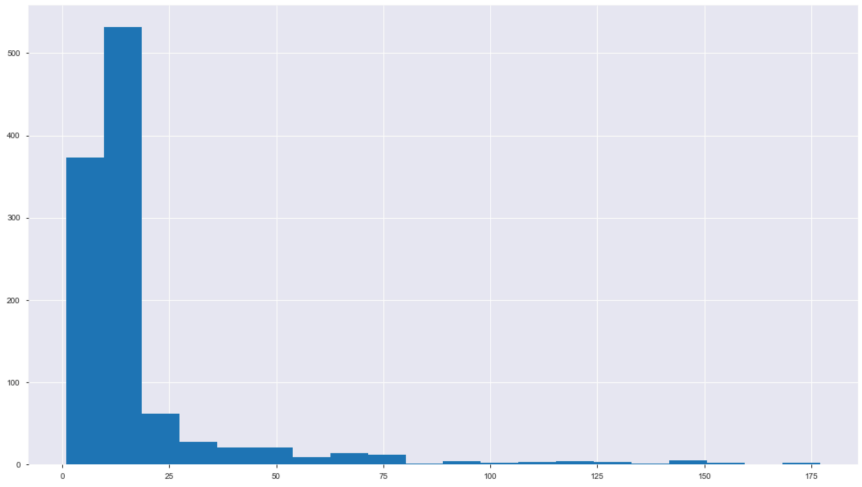


Figure 5.1: Age of Privacy Statement in Months, Full Dataset. The majority of statements are less than 2 years old, with the rest of the statements mostly being outliers.

and using smaller bins. As can be seen, the far majority was updated one year prior to this research took place, which was April 2018, indicating the large wave of updates to privacy policies made for the GDPR.

The far majority of the websites analyzed have headquarters in the United States. This is most likely due to the dataset analyzed, for which the Top-level domains (TLDs) have been filter to only show English websites. The selection contains *.com*, the US national TLD. The details of this variable are show in table 5.8.

Further inspecting the headquarters variable also highlighted an aspect of the dataset which was not found earlier. Because the privacy statements where gathered using Amazon Mechanical Turk, workers from different locations searched for privacy statements of different websites. As websites are able to recognize your location, some international websites showed a local version of the website, and also a local version of the privacy statement. This resulted in the local office also being identified as the headquarters. When looking for instance at entries with the headquarters India, of the 29 found, 8 where companies which in reality have US headquarters. India had this shortcoming in the highest percentage, other countries had less to zero errors. However, this is not necessarily a shortcoming; the fact that India was recognized as a headquarters means the privacy statement was edited specifically for that region. As the goal of the regional analysis is to assess whether privacy statements differ for users from different regions, it is only obvious to take these into account. We therefore assume that each privacy statement is written for the region where it is being viewed.

The majority of websites analyzed belong to the category *Technology & Computing*. This is most likely due to the nature of the dataset, as the most visited websites are analyzed. It is therefore more likely that the company behind the website has the website as its main platform through offering their service (e.g. Google, Facebook and YouTube). These types of companies belong to the 'Technology & Computing' category. Table 5.9 shows the most prevalent categories present.

Lastly, the 'Company Information' group also contain variables concerned with contact information which is provided in the privacy statement. Of all privacy statements, 75% provided an e-mail address and 62% provided a physical address to which letters could be sent. Lastly, 28% of the companies state that upon making (significant) alterations to the statement, the user will be updated via e-mail.

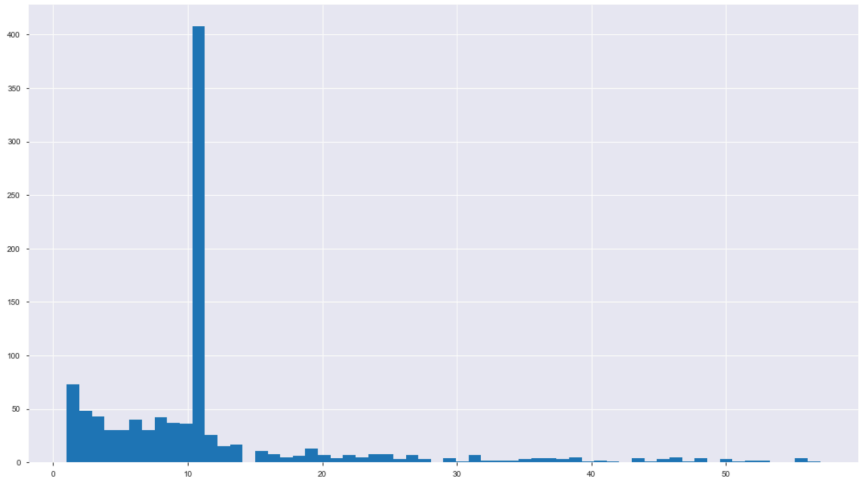


Figure 5.2: Age of Privacy Statement in Months, New Dataset. The figure shows the clear influence of GDPR, which obligated nearly all websites to update their privacy statements. The GDPR went into effect 11 months before the data was collected.

Table 5.8: Headquarters of companies behind websites, frequency table

Country	**Count**	**Frequency**
United States	1029	68.1%
None Found	230	15.2%
India	29	<0.1%
Canada	17	<0.1%
China	14	<0.1%
Switzerland	20	<0.1%
United Kingdom	19	<0.1%
France	14	<0.1%
Australia	11	<0.1%
Iceland	10	<0.1%
Other Values (26)	59	<0.1%

Table 5.9: Website Category, frequency table

Category	Count	Frequency
Technology & Computing	470	30.9%
News/Weather/Information	286	18.8%
Business	121	8.1%
Arts & entertainment	74	5.3%
Education	69	<0.1%
Hobbies & Interests	66	<0.1%
Non-Standard Content	63	<0.1%
Health & Fitness	49	<0.1%
Personal Finance	43	<0.1%
Travel	38	<0.1%
Other Values (16)	231	14.8%

5.2 Variable Analyses

In this section, a number of analyses are performed to test for what information can be extracted from the data. Firstly, a regional analysis is performed, combining statements from Europe, North America and Asia/Oceania, in section 5.2.2. This is followed by a category analysis in section 5.2.3, where each the average scores per category are compared to the average of the dataset. These first two analyses are mostly done to test the validity of the dataset, in order to see if results lie in line with expectations, and to see if other interesting results come up. Following these two analyses, a number of companies are highlighted separately in section 5.2.4, to assess if the accuracy of the data is suitable to compare companies on a company-level, to see if results lie within expectations. Finally, the variables in the dataset are used in an attempt to highlight good and bad privacy statements, and analyze in which aspects these statements differ.

5.2.1 Methodology

To assess the possible impact of different variables on privacy statements, an analysis is made of the effect of an independent categorical variable (in the coming two sections Region and Category) on multiple dependent variables, being boolean, numerical and ordinal. In principle, in situations like these Chi-squared tests are used to measure dependent boolean variables, and *Analysis Of Variance* (ANOVA) tests for dependent numerical variables (Miller, Jr., 1997). This section will first show the significant results for the analysis of numerical variables, followed by the results of the analysis of boolean variables. The full results of these tests can be found in Appendix B

Kruskal-Wallis Tests

When using ANOVA, in order to obtain trustworthy results the following assumptions must be met (from (Miller, Jr., 1997)):

- *Normality of residuals*
- *Homogeniety of variance of residuals*
- *Independent observations*

These conditions are more easily met if group sample sizes are similar, which is not the case for this dataset. Testing these assumptions is therefore critical in order to validate results. In this research, the Shapiro-Wilk test for non-normality, or in short the Shapiro-method is used in order to test for normality of residuals (Miller, Jr., 1997). This test can easily be applied to the dataset in Python, using the **Scipy** package, which

contains a built in section to apply the Shapiro-method to a dataset. After applying the Shapiro-method to every numerical variable every test resulted in statistically significant values (i.e. below 0.05), leading to the conclusion that the residuals of the data are not normally distributed. Therefore a non-parametric version of the ANOVA-test is applied, namely the Kruskal–Wallis H-test, in which case differences in medians are test instead of differences in means (Kruskal & Wallis, 1952). For this test, the first two assumptions defined above do not have to be met.

For each Kruskal-Wallis test two hypotheses are defined, namely:

H0: No statistically significant differences can be measured in the dependent variable given each region/category

H1: Statistically significant differences can be measured in the dependent variable given each region/category

Whereby the dependent variable depends on the numerical value measured on the y-axis.

Hypothesis one can be accepted if the test yields a p-value lower than alpha (for this research, the standard p-value of 0.05 is used). To calculate the H-values, the python package **Scipy** is used, which contains the built-in formula for the Kruskal-Wallis test. In the case that the test is rejected, i.e. populations of the groups differ significantly in rank, Dunn's test (Dunn, 1961) is used a post-hoc evaluation, to assess in which variables the difference is measured.

For both the regional and the categorical analyses, these hypotheses were defined and tested for each dependent variable. All the defined hypotheses are deliberately not noted in this section to ensure readability, as 61 different dependent variables are tested for both region and category. In stead, the following section presents the results of this analyses and some short implications of these results. Ultimately, testing all hypotheses will lead to the possibility of answering the working hypothesis which is defined in the beginning of each section.

5.2.2 Regional Analysis

In this section, the data is analyzed to assess differences between privacy statements from different regions. The regions are identified from the headquarters variable, the workings of which are specified in appendix B. As the headquarters variable identifies the country were the headquarters is located, this variable can be translated into regions. Based on the data, 3 different regions are identified, namely *North America, Europa and Asia/Oceania*. Furthermore, an additional group was identified for which the NLP-algorithm could not define a headquarters, which is defined as *None Found*. The largest group is North America with 1034 statements, followed by None Found with 210 statements, Europe with 107 statements and Asia/Oceania with 103 statements. The skewed size of the groups follows the earlier recognized limitation of this research, namely that all analyzed statements are written in English. Because of this, a natural bias occurs towards companies located in English-speaking countries, in this case mainly the United States. All tests performed are done without the *None Found* region, i.e. statements for which there was no specified headquarters.

The ultimate goal of this analysis is to find significant differences between regions which coincide with existing literature. If this can be done these findings can function as a form of external validation, and other new findings hold more value. This is done by applying the Kruskal-Wallis test, as explained in the previous section.

To accurately apply the Kruskal-Wallis test, firstly all entries where the country of origin cannot be extracted from the privacy statement are removed. When the region cannot be identified, this potentially has implications for the informativeness of the statement and not any relation to the actual region of the

Table 5.10: Derived regions of companies, frequency table

Region	Frequency
North America	1061
None Found	239
Europe	107
Asia/Oceania	103

statement. Therefore these entries cause noise in this analysis as their true regions are unknown, and are therefore removed from the regional analysis. The Kruskal-Wallis test is then applied to each numeric and ordinal variable for all entries which have a region, and the smaller subset of newer privacy statements. The eight numerical and ordinal variables which are measured are;

- Word Count
- Sentence Length
- Flesch Score
- Vagueness
- Age in Months
- Informativeness; the sum of all boolean coverage variables from (Wilson et al., 2016)
- Data Collection; the sum of all boolean data collection variables, from (Harkous et al., 2018)
- Data Sharing; the sum of all boolean data sharing variables, from (Harkous et al., 2018)

When applying the Kruskal-Wallis H-test and Dunn's test to these variables, many statistically significant differences can be found. The only metric which does not measure significant differences, is *monthsOld* when measured on the full dataset. On the new dataset, this variable does exhibit statistically significant differences. The following paragraphs will elaborate on a number of these results, only for the new dataset. For the full results, including the full dataset, see appendix B.

Figure 5.3 shows the boxplots of the word counts of the privacy statements, divided by each region. For this example, the test proved that privacy statements originating from companies within North America and Asia are significantly longer than those from Europe. The same results were yielded using the dataset of all statements. Where North America and Asia/Oceania had an average of just under 4000 words per privacy statement, this average was 2825 for the EU.

For the Flesh Reading-Ease (FRE) score, the new dataset indicates a significantly higher value for the European region, indicating statements from Europe are on average easier to read. Dunn's test reveals while North America and Europe score similar, Asia scores significantly lower. This can also be seen in figure 5.4.

The boxplot in image 5.5 shows the different distributions for the age of the privacy statements in the groups, for the new dataset. For Europe the interquartile region is the smallest, averaging around 12 months prior to the extraction of the data (May 2018, the month of the adoption of the GDPR). Also notable is the large number of outliers detected for North America, which falls in line with the plausible issue addressed in the beginning of this section, of the data containing relatively old privacy statements of North American companies due to the high presence of the '.com' TLD. Mostly, this test shows the effect that the GDPR has had on the European companies.

Applying the tests to the constructed ordinal variables (informativeness, data collection and data sharing) also yield statistically significant results. Where Asia/Oceania and Europe mention 6 of the 8 privacy statement topics (from Wilson et al. (2016)) on average, North American statements mention 7. When looking at data collection, privacy statements of all regions collect, when rounding, an average of 5 types

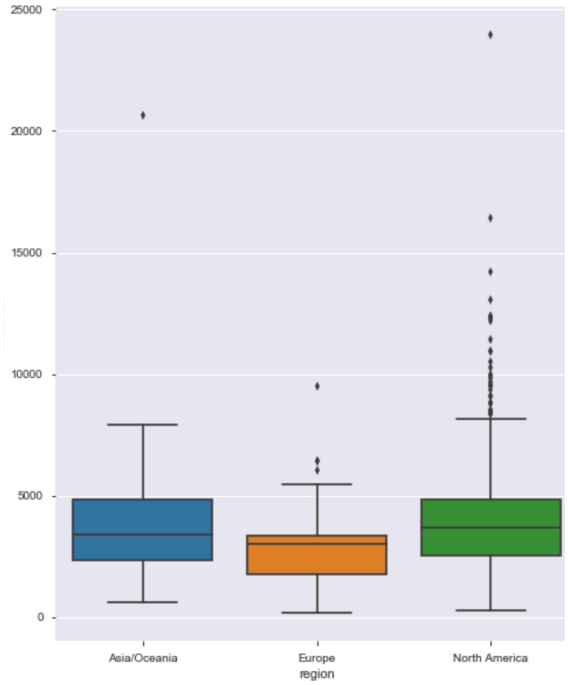


Figure 5.3: Boxplot of Word Count per Region. The figure shows European statements are signficantly shorter, and the influence of North America being overrepresented (more outliers).

of data from the user upon visiting. Nevertheless, test results show that North American companies collect more types of data on average than European and Asian companies. And lastly regarding data sharing, the average company (may) share 3 types of data from the user; this does not differ per region. The boxplot of data sharing can be seen in appendix B.

Chi2 contingency tests

To assess if a significant effect can be measured between the regions and the boolean variables in the dataset, a *chi-squared contingency* test is applied for each boolean variable. The chi-squared test can be used to assess differences in expected and actual values controlled for an independent categorical or ordinal variable (Beasley & Schumacher, 1995), in this case region. If the chi-squared test assesses significant difference, the test is repeated for each region with *Bonferroni Adjusted p-values*, in order to assess for which region and what differences can be measured (i.e. more or less actual values than expected).

In total, 34 of the 54 boolean variables showed significant differences in at least one of the regions. A full table of all results per variable is given in Appendix B. For each variable and each region an assessment is made if the variable is significantly higher or significantly lower than the average. Interestingly, nearly all significant variables were higher than the average for North America, and lower than the average for Asia/Oceania and Europe. In other words privacy statements from North America have mostly higher positive-rates for the boolean variables, and Europe and Asia/Oceania mostly lower. For elaboration on this, see figure 5.7.

In figure 5.7, the rows represent the variables for which a significant difference was found, the columns present the regions. The 3 left columns give the average frequency per region of each variable, the 3 columns on the right indicate if a higher than average significant, lower than average significance, or no significance (indicated with a hyphen) was found.

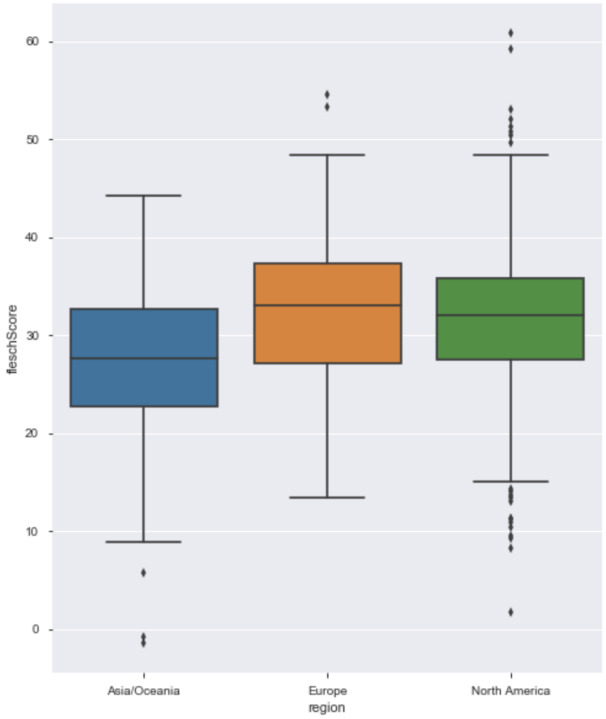


Figure 5.4: FRE Score per Region. As a high score indicates an easier to read text, European statements are on average the easiest to read.

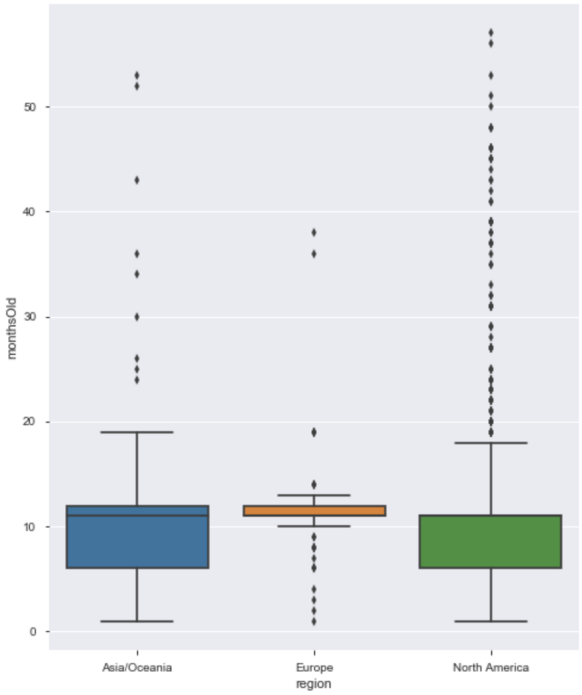


Figure 5.5: Age in months per region. The boxplot shows the clear influence of the GDPR; nearly all European statements are updated approximately twelve months prior to collecting the data, which is one month before the GDPR came into effect.

Although initially peculiar, in part these results are caused by the length of the North American privacy statements. As these statements are the longest, there is a higher likelihood that these statements mention a certain topic as they contain more content. To prove this, an additional test is done, with the 400 largest North American privacy statement removed from the dataset. This reduced the average word count of the North American statements from 3968 to 2536 (which is more similar to the European average word count of 2508). For this test, the variables measuring the presence of the topics 'legitimate interest', 'data protection officer' and 'data controller', i.e. words related to the GDPR, had a higher than average presence in the European statements, which is more in line with expectations (see Appendix B for the results).

As expected, European statements mention the data protection officer (DPO) significantly more often, as this is associated with GDPR. Also, European statements seem to cover significantly less of the 'standard' privacy statement topics, as defined by (Wilson et al., 2016). When regarding third parties, North American statements are nearly twice as likely to mention either Google or Facebook, and nearly seven times as likely to mention Amazon (although this percentage is already relatively small). North American statements are also twice as likely to mention combining their data with data of third parties. Lastly, the data suggests that North American statements collect and share significantly more data than European statements.

Conclusions

The analysis shows a large number of significant differences between groups. One of the most notable is the length of the privacy statements in words; the North American and Asian/Oceanic statements are on average more than 1000 words longer than then European privacy statements. European statements also tend to be significantly easier to read, and most of them are updated around the date that GDPR was adopted, May 2018. These results show the impact of the GDPR and are therefore in line with expectations; as the GDPR

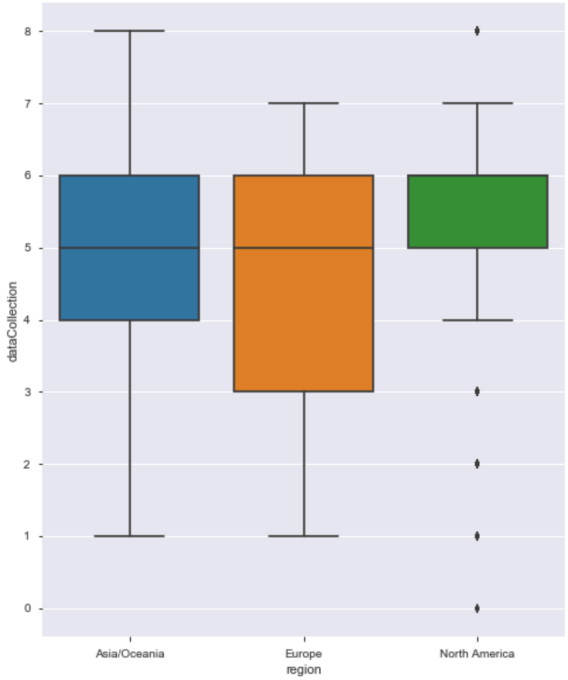


Figure 5.6: Data Collection per region. The figure shows than on average, European statements tend to mention collecting less different types of data than Asian or North American statements.

is a European regulation, the impact is expected to be strongest for statements written for European users. An important aspect of the GDPR is bringing clarity and reading ease to statements (see section 2.1.2). This is also reflected in the analysis, with European statements being the shortest and easiest to read, which strongly suggests this is due to the recent GDPR.

Due to the additional length of the North American statements combined with the way the algorithms assess the boolean variables (binary frequency, see section 2.2.1), the North American statements covered more of the topics for which was assessed, therefore yielding a higher than average frequency for nearly every variable tested (see figure 5.7). In other words, North American statements tend to cover topics more often, simply because they are longer. Only the variable assessing for the presence of the DPO, or Data Protection Officer, was higher for European statements, which is in line with expectations due to the fact that it is related to the GDPR.

An important notion to take into account when reviewing these results is that these are the results extracted from the privacy statements, i.e. derived from the information companies choose to put in their privacy statements. If full honesty and clarity is assumed from the companies, then the conclusions can for instance be made that North American companies partake significantly more in web tracking than European companies. However, it is not possible to currently make that assumption as honesty and clarity from companies cannot be tested for. Thus, these factors are important upon making conclusions. What can be concluded from these results is that North American privacy statements mention that they collect information to track users' web behavior more than European and Asian/Oceanic companies. This has implications for this research, which will be discussed in chapter 6.

	North America	Europe	Asia/Oceania	North America	Europe	Asia/Oceania
dpo	25.7%	46.2%	20.7%	-	higher	-
thirdParty	99.2%	93.8%	98.3%	higher	lower	-
choices	87.6%	72.3%	56.9%	higher	lower	lower
security	97.3%	87.7%	96.6%	higher	lower	-
specificAudiences	73.8%	49.2%	46.6%	higher	lower	lower
privacyShield	32.8%	13.8%	3.4%	higher	lower	lower
google	63.3%	38.5%	32.8%	higher	lower	lower
facebook	53.3%	33.8%	34.5%	higher	lower	lower
amazon	10.2%	1.5%	1.7%	higher	-	-
socialMedia	67.6%	53.8%	44.8%	higher	-	lower
daa	26.7%	4.6%	6.9%	higher	lower	lower
android	17.4%	3.1%	15.5%	-	lower	-
emailProvided	84.1%	87.7%	62.1%	-	-	lower
computerInformation	89.2%	80.0%	81.0%	higher	-	-
webTracking	56.3%	29.2%	44.8%	higher	lower	-
dataCombining	16.5%	7.7%	6.9%	higher	-	-
demographicInformation	76.7%	61.5%	75.9%	-	lower	-
contactInformationShare	46.1%	33.8%	34.5%	higher	-	-
personalInformationShare	86.7%	70.8%	89.7%	-	lower	-
generalShare	97.3%	90.8%	96.6%	-	lower	-
changeNotification	35.8%	15.4%	19.0%	higher	lower	-
addressProvided	77.1%	63.1%	53.4%	higher	-	lower

Figure 5.7: Variables average percentages left, Chi-squared results right

5.2.3 Category Analysis

In this section, the same types of analyses are performed for the independent variable website category. As with the previous section, this analysis serves partially as validation to test if the known effects of website category can also be measured in this research. If successful, it is also interesting to find out in what other aspects privacy statements differ. That privacy statements differ for certain website categories has been analyzed, but this research provides the possibility of analyzing for multiple variables on a large number of statements, leading to the possibility of making additional findings. Also,

Kruskal-Wallis tests

As in the previous section (section 5.2.2), the Kurskal-Wallis test is also applied for the category analysis. The applied Shapiro-method also showed that when taking the categories as independent variables, residuals are also not distributed normally, meaning the data is unfit for an ANOVA-test. Therefore the Kruskal-Wallis test is applied for the same 5 numerical and 3 ordinal variables as before, combined with the Dunn's test in case of significance to assess where the significant differences are. To ensure valid results are obtained, categories with too low frequencies are removed from the data. This results in a dataset with 1416 statements, and 15 different categories. Applying the same division to the new dataset provides the same 15 categories, with 975 values. However, as no significant differences can be measured for the age of statements between categories, the expected effect of privacy statement age is limited. After testing all variables for both the new and the full dataset, no differences where found; therefore, all tests for the category analysis are done with the full dataset, as the full dataset provides larger samples per category, leading to more valid results. Some difference were found between the full- and the new dataset for the chi-squared tests, these are provided in Appendix C.

In terms of word count and sentence length (word count shown in figure 5.8), the most notable category is Law, which has significantly shorter privacy statements and shorter average sentence lengths than other categories. Technology & computing, Personal Finance, and News, Weather & Information all have significantly shorter privacy statements, while Travel and Style & Fashion have significantly longer privacy statements. For average sentence length the differences are less, with Sports having a significantly long sentence length, and Law significantly shorter sentence length.

For the Flesch Reading-Ease (FRE) score, 2 categories are notable, namely Sports with significantly more complex statements, and society with significantly less complex statements. With an average FRE score of 28.7, the average Sports statement can be understood by university graduates and is rated as 'Very Confusing', while with an average FRE score of 34.9, the average Society statement is understood by college students and rated as 'difficult'. No other large significant differences are found within the categories. For the boxplot of FRE-score per category, see Appendix C.

Significant differences are also found for the vagueness metrics, which assess the usage of vague words within a statement. The significantly most vague statements are those from the categories 'Arts & Entertainment' and 'Health & Fitness', while the least vague statements are those of the categories 'Education', 'Business' and 'Technology & Computing'. As mentioned earlier, no significant differences are recognized for the age of privacy statements.

The informativeness metric, which assesses if the the 8 categories defined by (Wilson et al., 2016) are mentioned or not in a statement, also yield significant differences. As the categories are deemed to be basic categories which every statements should assess, a higher score on this metric can be read as a more informative privacy statement, which is desirable as it indicates the transparency of the company. Because of this, the Dunn's test and the boxplot (see figure 5.9) reveals a number of more varying- and constantly informative statements. From the test, the categories 'News, Weather & Information', 'Technology & Computing', 'Law' and 'Personal Finance' are recognized as the categories where informativeness varies most. The categories 'Health & Fitness', 'Business', 'Hobbies & Interests' and 'Arts & Entertainment' have on average the most informative statements.

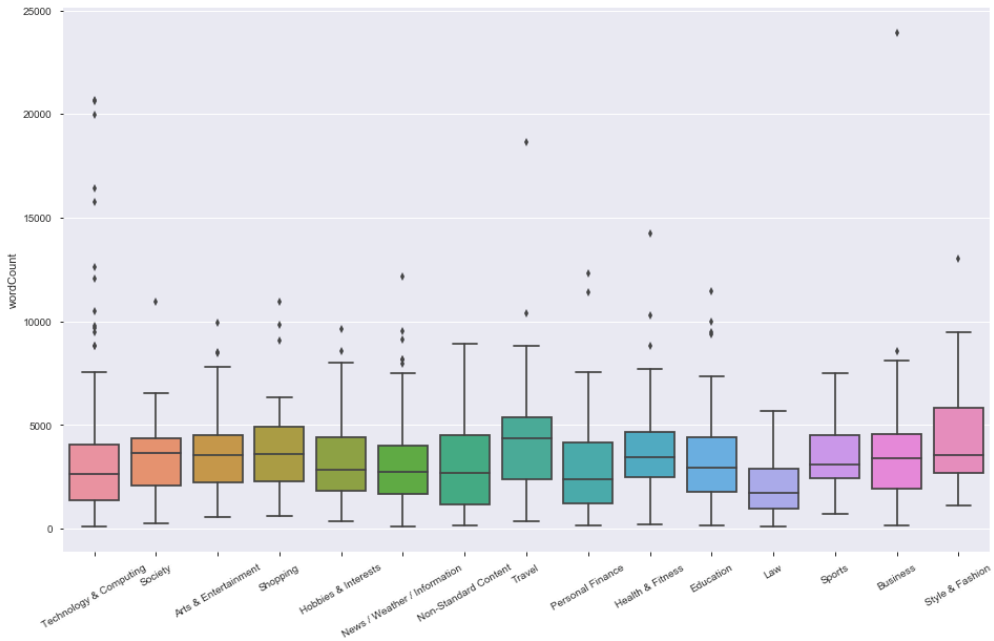


Figure 5.8: Word count per category

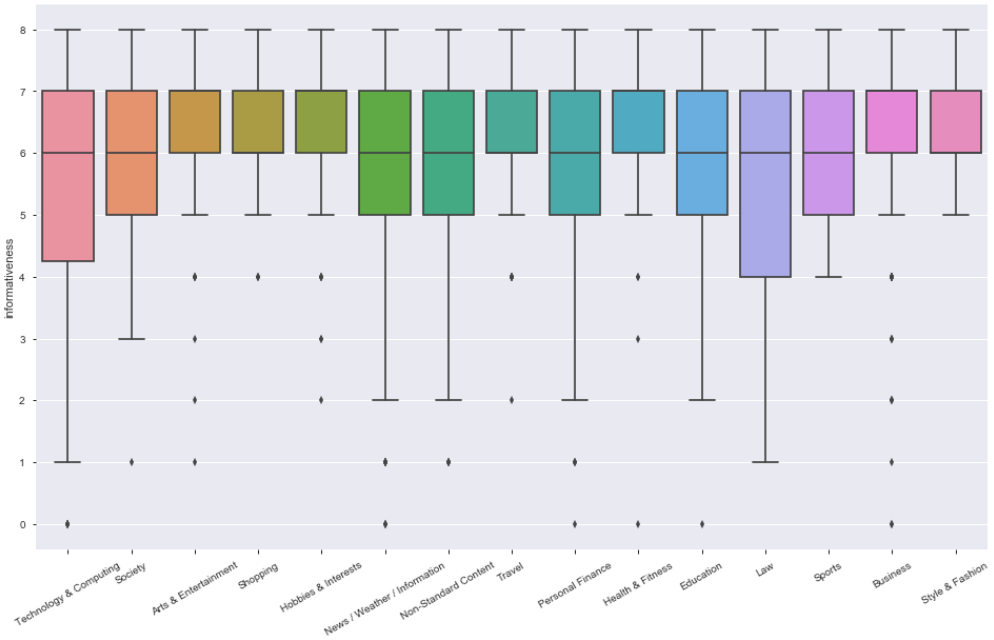


Figure 5.9: Informativeness per category. Although average values per category are high, interestingly some categories (such as Technology and Law) have many low entries.

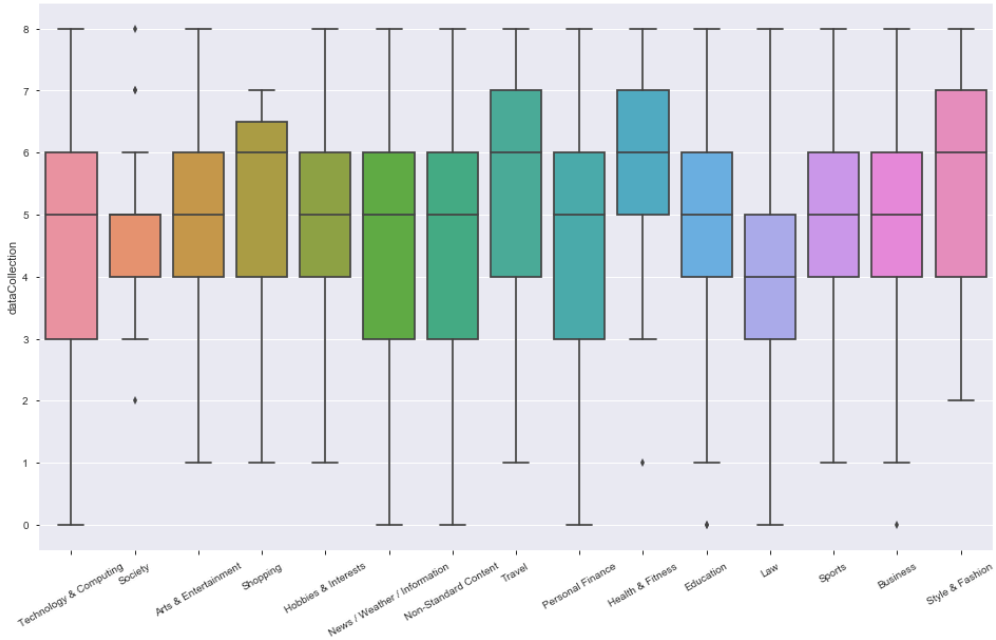


Figure 5.10: Data Collection per category, indicates the number of data types which are possibly collected from the user.

Testing the categories for the number of types of data collected also yields significant results. Categories 'Health & Fitness', 'Travel' and 'Style & Fashion' collect more data than average; especially 'Health & Fitness' stands out, as the only category which collects a rounded average of 6 possible types of data collection. Law collects the least types of data, on averaging 4 types of data (see figure 5.10).

Lastly, significant differences can be found between categories for sharing of data as well. While 'Law', 'Technology & Computing', and 'Education' on average share the least types of data (all averaging around 2 types), 'Travel', 'Style & Fashion' and 'Shopping' share the most types of data (all averaging around 3 types of data). The results can be seen in figure 5.11.

Chi-squared contingency tests

As in the previous section, chi-squared contingency tests are applied to assess if groups of the independent variable (in this case categories) differ significantly from the average of the dataset. If significant differences are found for a variable, the test is repeated with Bonferroni adjusted p-values to validate the significancy, assess between which variables significant differences are found and if these differences are higher or lower than expected. As the resulting table of this analysis is quite large, it is shown in Appendix B. This section will provide a short summary of the results per category.

'News, Weather & Information' (NWI) has the most significant differences from the average measurement. The clearest trend that can be identified from the results is that variables related to the GDPR are on significantly lower than the average, indicating these statements focus less on GDPR. Furthermore, NWI statements have a lower chance of discussing the topic of data retention or of using Privacy Shield, but a higher chance of cooperating with *DoubleClick*, Google's adservice.

'Technology & Computing' is also a notable category with many differences from the average. For data

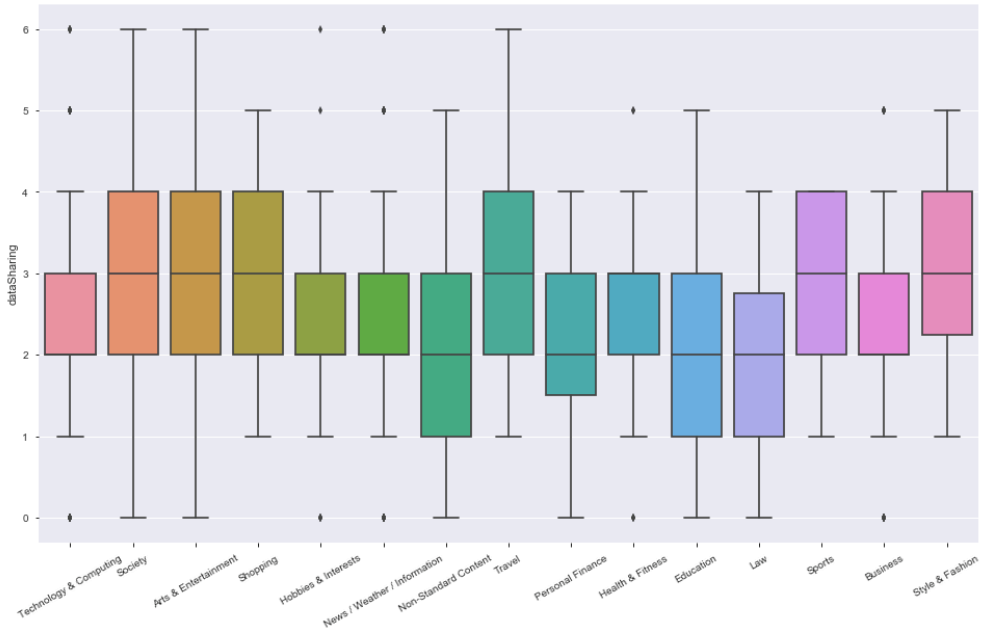


Figure 5.11: Data Sharing per category, indicates the number of data types which are possibly shared with third parties.

collection and -sharing, the category averages lower on 3 and 2 different variables respectively, indicating this category collects and shares significantly less data than other categories.

Other categories that exhibit significant differences have this for less variables than the first two mentioned. An interesting finding is the category 'Shopping', which scores significantly higher for sharing of financial information. The categories 'Business' and 'Travel' score higher than average for variables related to GDPR. The categories 'Sports', 'News, Weather & Information', and 'Arts & Entertainment' score higher on variables related to third parties, indicating these websites are more likely to work together and share data with third parties.

Upon comparing these results to the results achieved from the new dataset, similar results where achieved but with less significance, indicating that this variables likely does not have a large effect on the analysis, as discussed earlier in this section. These results are discussed in Appendix B.

Third Party Overview

A last assessment which is done for the category analysis is to compare the average presence of third-parties per category. Although this is partly done in the previous section, this section simply provides the descriptive statistics to see which category mentions which third party the most, even if this is not a significant difference from the average metric. The full table can be seen in figure 5.12, where blue cells indicate a high presence and yellow cells indicate a low presence. A selection was made of the most interesting third parties based on the results, the full table can be seen in Appendix B.

Some categories yield interesting figures; business, for instance, has the highest presence of Privacy Shield by some margin. Health & Fitness also yields interesting results, mentioning, Google, Amazon, Double Click and social media in general very frequently. Style & Fashion have the highest presence of Facebook and

Category	Privacy Shield	Google	Facebook	Amazon	Microsoft	DoubleClick	Apple	Social Media
Technology & Computing	27%	52%	40%	9%	7%	8%	8%	47%
News / Weather / Information	12%	57%	42%	8%	10%	17%	14%	57%
Business	41%	58%	46%	7%	7%	9%	7%	56%
Arts & Entertainment	21%	64%	56%	16%	12%	14%	25%	63%
Education	22%	49%	32%	7%	9%	10%	10%	44%
Hobbies & Interests	32%	55%	52%	6%	8%	11%	15%	67%
Non-Standard Content	18%	49%	43%	8%	3%	5%	16%	40%
Health & Fitness	16%	67%	43%	12%	8%	14%	14%	71%
Personal Finance	12%	51%	30%	0%	12%	5%	7%	47%
Travel	24%	49%	49%	0%	0%	8%	11%	70%
Society	30%	49%	52%	6%	6%	9%	3%	49%
Shopping	20%	37%	43%	3%	0%	7%	10%	60%
Sports	11%	57%	43%	4%	25%	4%	29%	64%
Law	12%	42%	27%	0%	4%	19%	4%	46%
Style & Fashion	18%	46%	64%	5%	9%	9%	14%	77%

Figure 5.12: Third party presence per category. Blue values indicate a third party has a high presence in that category, yellow indicates a low presence.

social media, while Sports tend to mention Microsoft and Apple the most. All these connections indicate that some categories tend to work together with these companies more than others.

An important note here is that these results are simply descriptive, but do not present any statistical significance; they provide an overview of what is measured, but not if these are statistically significant from the average. These measurements where done in the previous section, section 5.2.3. The results from the previous section can however clearly be seen in this table.

Conclusions

Using the achieved dataset, many significant difference can be found between the privacy statements of websites from different categories. For the conclusions the categories which had the most differences are shortly discussed.

Technology & computing is a category with many significant results; the average Technology & computing statement is often short and concise (not vague), and tends to mention collecting and sharing data less. Travel websites are often the opposite, having relatively long privacy statements, but collecting and sharing many types of data. This would seem in line with expectations; technology & computing websites can be assumed to have knowledge on technology and therefore data privacy, which leads them to only collecting the necessary data, and making sure to notify this in short, concise manner. Travel websites need to collect and share data from their users in order to complete bookings for flights, hotels or car rentals.

Another notable category is News, Weather & Information (NWI). NWI-websites tend to have shorter and less informative privacy statements, which cover less topics related to GDPR. Double Click, Google's ad service, is mention significantly more often in these statements. This lies in expectations with existing literature on the online presence of NWI websites (Libert & Nielsen, 2018).

Law is also a notable category: while not having a low average in informativeness, these statements tend to be short and websites tend to collect and share less data than the average website. However, Law websites do mention Double Click the most often of any type of website. Business websites are somewhat similar, being informative and concise, and using Privacy Shield the most of all categories.

Lastly, Health & Fitness is also a notable category. As health data concerns a special category of data (European Union, 2016), these websites would need to be careful if data is collected. While being some of the most informative statements, these websites also collect the most types of data from the user, and with very high frequency mention Google and/or some kind of social media. These results lead to further questions; although some of these results lie in line with expectations (Technology & Computing, Travel and NWI) others, namely for the category Health & Fitness, could be interesting for further research. This is further elaborated on in section 7.3.

Ultimately, multiple measurable differences were found, some of which can be found in existing literature. This validates the working of the variables and gives possibilities for interesting findings in the data and further future work.

5.2.4 Company-Level Analysis

When discussing data privacy in the current age, a lot of interest is put into the big-5 tech companies. These 5 companies (Google, Facebook, Apple, Amazon and Microsoft) are by and large the dominate players in the internet market with some of the largest data practices (Dolata, 2017). Most of these companies are the companies which Zuboff (2015) describes as the pioneers of the surveillance capitalists, i.e. companies which aim to collect as much data as possible from their users. Google and Facebook have already been subject to large data privacy scandals (Isaak & Hanna, 2018; Zuboff, 2019), bringing their actions into attention, and making their users question their attitudes towards privacy. On the other hand, Apple profiles itself as a company that has a high regard for privacy, as their core business model is selling products and not advertisements (Graham, 2018; Zuboff, 2019).

If we accept these statements as truths, then the conclusion can be made that of all these data intensive companies, Google and Facebook, and to a slightly lesser extent Amazon and Microsoft (Zuboff, 2019) are surveillance capitalists, but Apple is not (at least to a far lesser extent (Graham, 2018)). Each company therefore holds differing values for a customer's privacy. While some attempt to safeguard privacy, others can only exist when massive amounts of data are collecting, because that is how their business model works. Using this information, this section will see if the privacy statements of these companies exhibit differences which are in line with the business models these companies have, and also in line with the expectations of the behaviors of surveillance capitalists.

To guide this analysis a working hypothesis can be defined. To recapitulate, a working hypothesis is a hypothesis which is provisionally adopted, in order to explain a still unproven relationship, in order to guide research (Shields, 2003). For this section, the following working hypothesis is tested:

Working Hypothesis: ***The privacy statement of a surveillance capitalist differs on certain variables from a privacy statement of a similar company which is not surveillance capitalist.***

To test the working hypothesis, a number of expectations can be defined and tested by simply comparing the expectations to the variables. This is a very simple method, but can be used a first indication to if clearly observable differences exist between the companies. In this case, they are formed based on what surveillance capitalists would score on the variables extracted from the privacy statements. The variables selected for this analysis are either single variables which are not extracted via NLP, or composite variables consisting of multiple NLP-variables. This is done to ensure accuracy, as combining single NLP-variables on a single statement basis has a higher chance of errors; this is further discussed in the Discussion (chapter 6. The variables compared in this research are therefore: *Word Count, Sentence Length, FRE-score and Flesch Level, Vagueness, Informativeness, Data Collection* and *Data Sharing*. The last three of these are composite variables, consisting of all variables related to that topic (for more info on these variables, read chapter 4.

Table 5.11: Readability Metrics for Big-5 Tech companies

	Google	Facebook	Amazon	Microsoft	Apple
Word Count	7153	4300	2639	2791	4009
Sentence Length	25,92	24,86	23,77	25,14	26,91
FRE score	39,82	42,03	34,44	33,99	29,38
Flesch Level	difficult	difficult	difficult	difficult	very confusing
Vagueness	20,55	15,12	10,99	8,24	20,45

As discussed in section 2.1.3, a surveillance capitalist attempts to collect as much data as possible from their users as that is part of their business model. It also actively tries to hide doing so, as informing consumers could lead to less data, and therefore less revenue. From these statements, one can expect a surveillance capitalist's privacy statement to be long, vague and complex: clearly stating what is done with data might lead to a reduction in the amount of data. A first expectation defined is therefore:

Expectation 1: *The privacy statement of a surveillance capitalist is longer, harder to read, and vaguer than that of a similar company which is not surveillance capitalist.*

This working hypothesis can be tested by assessing the first five variables as defined in the previous paragraph for the big-5 tech companies. Second, the surveillance capitalist is expected to collect and share many types of data, as that is part of their business model (also described in section 2.1.3). The second working hypothesis can therefore be defined as:

Expectation 2: *Surveillance capitalists collect and share more types of data than similar companies who are not surveillance capitalist.*

Firstly, table 5.11 shows an overview of the readability scores for each company. On average, Apple's statement scores the worst for readability, with long sentence length, lowest FRE score and therefore highest Flesch reading-level, and highest amount of vagueness. Facebook and Google have the easiest to read privacy statements. Microsoft and Amazon's privacy statements are the shortest, containing the least amount of vague words.

Next, the 3 combined ordinal variables of informativeness, data collection and data sharing are compared. These variables indicate the number of privacy statement topics addressed, the number of types of data collected, and the number of types of data shared respectively. The results are show in table 5.12. The table shows that all statements are similar in informativeness, with Facebook and Amazon lacking slightly. Interestingly, Apple clearly states mentioning the most types of data, followed by Google, Microsoft and Amazon, and Facebook states collecting the least types of data. In regards to sharing, Apple states sharing the most different types of data, followed by Google, Facebook, Microsoft and Amazon respectively.

Table 5.12: Informativeness, Data Collection and Data Sharing for the Big 5 Tech companies

	Google	Facebook	Amazon	Microsoft	Apple
Data Collection	6	4	5	5	7
Data Sharing	4	3	1	2	5

Relating back to the defined expectations, we see they are both assumed to be incorrect based on this analysis. While the surveillance capitalists privacy statements were longer, their statements were also easier to read and less vague. Furthermore, their statements indicated that they shared and collected less information than

a company with a high privacy commitment. Because of this, the working hypothesis can not be proven using this method.

Conclusions

Interestingly, results of this analysis are more opposite of the expectations: as described by Zuboff (2015), surveillance capitalists are companies which thrive off of collecting massive amounts of data, and either using or selling these for profit. As such, Google and Facebook would be expected to stand out for data collection and data sharing. Furthermore, Zuboff (2019) also describes that companies aim to be secretive about these operations. As of such, one could expect a surveillance capitalist's privacy statement to be harder to read, and more vague. Both of these are not the case. In fact, Apple, the company which should stand out for it's attitude towards privacy, has the most complex and the second most vague statement. From table 5.11, no clear indicators can be seen which point out that the surveillance capitalists have differing privacy statements from Apple. If anything, Apple's statement is less readable then those of surveillance capitalists. The same opposite results are yielded for data collection and -sharing. Apple collects and shares the most different types of data according to the companies privacy statements, followed second in both categories by Google. As surveillance capitalists aim to collect as many different types as data as possible (Zuboff, 2015), this is also opposite of the expectations.

However, these results do impede the validity of this research. On the contrary, these results could provide further insights on the state of the privacy statements of these companies. In this research the logic of surveillance capitalism is accepted to be true, and as such Google and Facebook are highly likely to collect more different types of data than Apple. As 7 types of data were found for Apple (and the algorithms are more likely to have false negatives than false positives, see section 4), the expected amount of collected types of data would be higher for Google and Facebook. If this hypothesis is true, then it is possible that the statements are lacking information, or that the collection and certain types of data-sharing or -collection are not stated clearly enough, or not stated at all. (resulting in a false negative for the algorithms). Summarized, when analyzing the data of statements on an individual level, in this case, a number of values differentiated from the expected values. Possibly, this could indicate that surveillance capitalists (i.e. Google and Facebook) are less clear about what data they collect and share in their privacy statements, resulting in lower results for this analysis. However, proving that statement would require additional research.

5.2.5 Statement Quality Analysis

In this section, an attempt is made to identify statements which are deemed to be good, and statements that are deemed to be bad. To do this, first some assumptions must be made on what defines a 'good' and a 'bad' statement. As described in section 2.1.2, the goal of a privacy statement is *"to inform users in which ways data is gathered, used, shared and managed by a company's services or products"*. The GDPR also emphasizes this in article 12, stating privacy statements must be provided *"in a concise, transparent, intelligible and easily accessible form, using clear and plain language"* (European Union, 2016).

Translating these two quotes to the variables measured in this research, an informative statement is informative if it covers all necessary topics for a privacy statement. Wilson et al. (Wilson et al., 2016) defined these necessary topics, and this research has ennumerated them in the variable 'informativeness'. Therefore, if a website's privacy statement has a high informativeness score, the better coverage the privacy statement has of all topics. Second, as the GDPR states, it is important to write in 'clear and plain language'. The applied FRE-score assesses the readability of texts, and therefor gives a clear indication of the 'clarity and plainness' of the text. Lastly, it can be said that given the informativeness and clarity of the text, a privacy statement must not be too long. It is important to emphasize that length must only be short given informativeness and reading ease; once a privacy statement lacks informativeness it can become too short, therefore meaning there is a right length for privacy statements (also defended by Hintze (2017)). But a too long privacy

Figure 5.13: Scatterplot of reading ease, word length and informativeness

statement decreases it's potential of informing the user (Cranor, 2012), thus making it important to find the right balance.

These metrics are combined and plotted in the scatter plot in figure 5.13, with word count on the x-axis, FRE-score on the y-axis, and the color of the dots indicating informativeness (where a lighter color indicates a more informative statement). A clear correlation can be seen between informativeness and word count (correlation of 0.53); as statements become longer, they tend to cover more topics. It also indicates a certain 'sweet spot', where the optimal privacy statements are; this sweet spot is found where informativeness is maximum, FRE-score is as high as possible, and length is as short as possible. These statements are able to cover all important topics in clear language, as short as possible, and are therefore deemed to be most effective at informing their user.

To gain further insights into the effects of these metrics and if there are trends within these groups, a division is made between best- and worst practices of privacy statements based on these metrics, and a check is done for which variables the differences are largest.

First, best- and worst practices of privacy statements are defined using the three metrics of word count, FRE-score and informativeness. For each statement, variables are made which assess if they are below average on word count, above average on FRE-score and above average on informativeness. If all three are the case, this is defined as a best practice. Of the dataset of 1508 statements, 234 of these best practices where identified. An initial thought would be to apply the same rules in an opposite way, in order to obtain the worst practices. However, applying all three rules results in only 31 statements. This is mostly due to the correlation between word count and informativeness; i.e. there are not many privacy statements which are above average in word length, and below average in informativeness. To improve the validity of the

results, the worst cases are only defined based on FRE-score and informativeness. This still fully separates the groups, but enlarges the worst-case group to 194 statements. The division is visualized in figure 5.14.

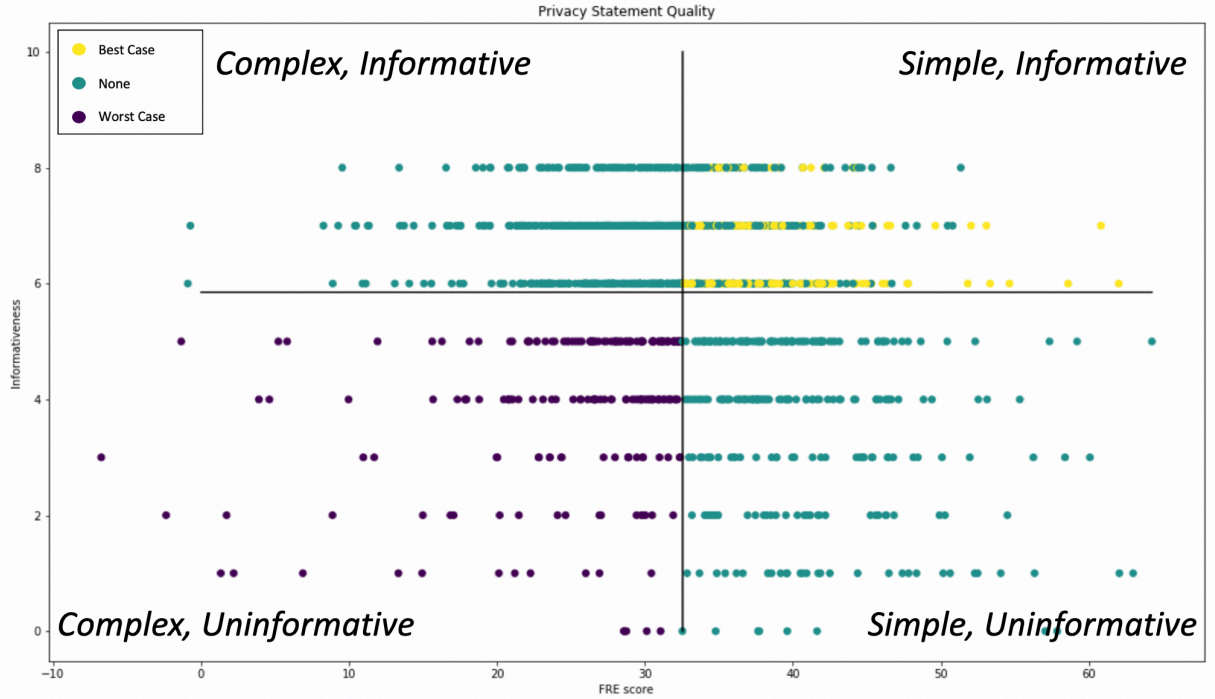


Figure 5.14: Best- and worst cases, plotted for FRE-score and Informativeness

Figure 5.14 shows another scatter plot of all statements, but with FRE-score on the x-axis and informativeness on the y-axis. Lines are added which indicate the mean values of both axes lie. As can be seen from the figure, because best cases also need to be below average on word count, not all statements in the top right section are best cases. As this rule does not apply for the worst-cases, all statements in the bottom-left are worst cases.

The next step is to assess if there are variables for which the two groups have significant differences. To do this, the same methods are used as in the regional analysis (section 5.2.2) and the categorical analysis (section 5.2.3). Firstly the numerical and ordinal variables are assessed using the Kruskal-Wallis H-test (excluding FRE-score and informativeness as these have contributed to the separation of these groups).

After performing the tests, the best-case statements are on average longer (2280 words average, compared to 1880 average for the worst cases), have shorter sentences (5 words per sentence shorter) and are slightly more vague. Even though the best cases are selected for having below average word count, the worst-cases are on average shorter as they are selected on low informativeness, which correlates with word count. This is in line with the argument made by Hintze (2017), that privacy statements need a specific length in order to be informative enough. Interestingly, the best-case statements also collect and share significantly more data than the worst-case statements. These findings are interesting, as they indicate that companies that put much effort into their privacy statement do not necessarily limit the amount of data which is collected

and shared.

Lastly, the boolean variables are analyzed for significant differences between the groups. The same Chi-squared contingency tests with Bonferroni-corrections are applied. The result of significantly differing variables with their corresponding frequencies are show in table 5.13.

Table 5.13: Significant differences for boolean variables. Left: effect difference. Right: True-frequency within the group

	Worst-case	**Best-case**	**Worst-case**	**Best-case**
Vital Interest	higher	lower	12.9%	2.6%
Choices	lower	higher	16.1%	88.9%
Specific Audiences	lower	higher	16.1%	77.4%
Privacy Shield	lower	higher	0.0%	21.4%
Data Retention	lower	higher	48.4%	75.2%
Microsoft	higher	lower	22.6%	4.3%
Apple	higher	lower	19.4%	3.4%
General Share	lower	higher	87.1%	99.1%

Some variables are not surprising: the variables *choices*, *Specific Audiences*, *Data Retention* and *General Share* are four of the eight boolean variables which make up informativeness. As these groups are divided on informativeness, a number of these variables where to be expected in the results. It is interesting to note that these variables are the variables of informativeness which mostly define a good and a bad statement.

Some results are surprising. For instance, the worst-case statements have a more than 5-times higher likelihood of mentioning vital interest, one of the legal bases for collecting data from GDPR. The worst-case statements also have a far higher likelihood of mentioning big-tech companies Microsoft and Apple. Lastly, the worst-case statements never mention Privacy Shield, indicating that the Privacy Shield framework has a positive effect on the state of a company's privacy statement.

Conclusions

Based on the literature, three variables are used to identify the quality of a privacy statement: informativeness (Wilson et al., 2016), FRE-score and word count (European Union, 2016). Based on these three variables, a number of best- and worst case statements are defined, resulting in 234 and 194 statements respectively. These statements where analyzed to find the largest differences between the groups.

The average best-case statement was able to be relatively informative with 2280 words, and used shorter sentences on average. The best cases also mentioned collecting and sharing more data than the worst cases. Furthermore, a number of indicators where recognized which differentiate the two groups. Best-case statements are far more likely to discuss user choices, policies for specific audiences, how long data is retained, and the fact that data is shared. Best-case scenarios are also far more likely to mention the use of the Privacy Shield framework. On the other hand, worst-case statements (interestingly) have a far higher chance of mentioning vital interest, and the companies Microsoft and Apple.

Apart from the fact that these results form useful signals for the quality of a privacy statement, an interesting conclusion can be derived from the results regarding data collection and data sharing. The best-case companies, which are selected on informativeness, stated that they collected and shared data significantly more than the worst-case companies. The chances that in reality the websites related to these statements actually collect and share more data, however, is small, as the companies where selected on the informativeness on the privacy statement, not any other factor which would have a significant impact on the collection and sharing of data (for instance, website category, as proven in section 5.2.3). A more realistic probability could be that

the websites belonging to these best- and the worst-case statements collect and share a similar amount of data, and that the issue is with how the privacy statement is written. In this case, the 'worst-case websites' simply do not clearly elaborate on all the types of data they share and collect, further emphasizes that the statement is poorly written. If this is true, these metrics could provide a useful tool for governments and other regulatory bodies, to advise companies to try and update their statement.

5.2.6 Clustering Analysis

In the previous section the data was analyzed to find differences between groups of privacy statements, to assess what effect these groups had on the data. A second step performed in this analysis is looking for similarities in the data, and more specifically clusters. One of the main motivations of this research was surveillance capitalism (see section 2.1.3). Zuboff (2015) recognized and explained a new form of capitalism, where companies would want to maximize the amount of data they collect from users. Companies that for instance simply use their website to share information on what products are sold would not need to collect such significant amounts of data. Because of the database created for this research, as assessment can be done to see if companies with similar privacy statements can be grouped, and if so, what these groups indicate. A first step in this process is to find if there are natural groups in the data, i.e. clusters. In reality, if groups were found in the data, these groups would represent groups of companies with similar aspects of privacy statements, which possibly entails similar ways of handling data, depending on the clusters.

To assess the presence of these natural groups, multiple clustering methods and validity metrics were tested on multiple subsets of the data. Although this section will not go into full detail on these methods, they are stated in Appendix A. In short, the methods K-means (Provost & Fawcett, 2013) and DBSCAN (Ester, Kriegel, Sander, & Xu, 1996) where used to assess clusters within the numerical and ordinal variables, using euclidean distance (Provost & Fawcett, 2013) as a distance metric. The methods K-medoids (Huang, 1998) and DBSCAN where used to assess for clusters all numerical and boolean variables, using Gower's distance (Gower, 1971) as a distance metric. The validity of the clusters was assessed using the silhouette score (Rousseeuw, 1987) and the S_dbw index (Halkidi, Batistakis, & Vazirgiannis, 2002). A number of tests were also performed on smaller subsets of the data, consisting of variables which were expected to have a larger impact in clusters.

Although a number of the tests showed slightly promising results for the validity scores indicating some form of cluster structure, manual inspection of the data showed no actual clusters. Furthermore, the most promising clusters identified had no directly logical relationship to reality, and were not in line with any expectations or known literature. After the rigorous testing, it is concluded that within the data extracted by this research from the privacy statements, no clear clusters can be found.

Although possible that the data does not exhibit clusters, there is also the possibility that the clusters could not be identified due to the types of variables extracted from the privacy statements. Traditionally, clusters are best found in datasets which mainly contain numerical variables (Provost & Fawcett, 2013). In this case, the data is a large, mixed dataset. To slightly improve the situation, a large number of the boolean variables where combined into three ordinal variables, namely informativeness, data collection and data sharing, which could potentially improve this situation. However, ordinal variables are not as optimal as numerical variables for finding clusters (Provost & Fawcett, 2013; Huang, 1998). When using Gower's distance, methods like DBSCAN and K-prototypes are able to detect clusters within mixed data, but even so, clusters are far less likely to occur in mixed data. Because of this, only if very apparent clusters would exist in the data, they could have been identified. This all therefore creates the possibility that potentially clusters do exist in privacy statements, but that they could not be identified because the research identified the wrong variables for clustering. Further elaboration on this is done in section 6.

If, on the other hand, it is accepted that clusters do not exist in the data, this can have a number of implications. Firstly, if the text in a privacy statement is accepted as the complete and accurate representation data practice of a company, then all companies handle data in slightly different ways, and no 'standard practices'

Table 5.14: Sentence Repitition, Full Database

Repeated	**1**	**2**	**3**	**4**	**5**	**>5**	**>10**	**>15**	**>20**
Frequency	146657	17540	5105	2835	1837	1264	99	33	13
Percentage	100,00%	11,96%	3,48%	1,93%	1,25%	0,86%	0,07%	0,02%	0,01%

exist for companies. A more likely scenario however is that the privacy statement does not contain enough detail to be a full representation of the data practices of a company, but forms more of an indication. The dataset used is able to extract a part of this indication, which is too limited to be able to cluster companies based on their data practices.

If clusters would have been found, these could have indicated global dependencies between variables extracted. For instance, that companies either mention many of the GDPR bases, or mention none at all. In reality, the data showed that this was far more nuanced, and no clear structures or 'dependencies' like these exist. Based on the data it would therefore seem that every company creates a privacy statement and handles the global topics of these privacy statements in their own, separate ways.

5.3 Similarity

An additional analysis is performed to test for similarity, which entails to what degree identical sentences are used between privacy statements. Finding out to what degree sentences are identical between statements can not only indicate if statements are similar or not, but also of statements copy certain parts of text. This analysis is done on two levels, namely for all statements combined on a sentence level, and on a statement level.

5.3.1 Sentence Level Analysis

For the sentence level analysis, all texts of the statements are combined to create a frequency analysis of each sentence (i.e. how often is each sentence stated). The data is pre-processed by separating each sentence using the sent_tokenize option from the **NLTK** package. Each sentence is also filtered to contain at least 6 words: a cut off point by manually reviewing the firstly yielded results, which contained many titles of paragraphs or hyperlinked sentences (for instance, *Read more here*, *Cookies* and *What information do we collect?*). This resulted in a list of sentences and their corresponding frequencies throughout the full database.

In total, 146657 different sentences were identified (longer than 5 words), of which 17540, or 12%, were repeated at least once. Of the repeated sentences the median frequency is 2, indicating most of the sentences are used twice. These sentences are likely to originate from two websites run by the same parent organization, which chooses to use the same language throughout the statement. After manual check, this was found to be the most frequent reason for identical sentences. A full overview of the frequencies of sentences can be found in table 5.14.

The more interesting sentences are those with the highest frequency, which are shown in table 5.15. When analyzing these sentences, roughly three groups can be defined: Firstly the informative/title group, which contains short, generic informative sentences ort titles (i.e. *Place of processing: united states - privacy policy.*, *How do we use your information?* and *In this case, we will notify you and keep you updated*). Second, a number of sentences can be attributed to a privacy organization, possibly implying the organization wrote the sentence and advised its customers or users to use the sentence (i.e. *if there is any conflict between the terms in this privacy policy and the privacy shield principles, the privacy shield principles shall govern* and *if you have an unresolved privacy or data use concern that we have not addressed satisfactorily,*

Table 5.15: Top 20 most-used sentences, with their corresponding frequency

Frequency	**Sentence**
40	place of processing: united states – privacy policy.
35	we consider this use to be proportionate and will not be prejudicial or detrimental to you.
34	we encourage you to periodically review this page for the latest information on our privacy practices.
34	if there is any conflict between the terms in this privacy policy and the privacy shield principles, the privacy shield principles shall govern.
31	if you have an unresolved privacy or data use concern that we have not addressed satisfactorily, please contact our u.s.-based third party dispute resolution provider (free of charge) at https://feedback-form.truste.com/watchdog/request.
30	your interactions with these features are governed by the privacy policy of the company providing it.
28	to learn more about the privacy shield program, and to view our certification, please visit https://www.privacyshield.gov/.
26	personal data collected: cookies; usage data.
25	your privacy is important to us.
23	these features may collect your ip address, which page you are visiting on our site, and may set a cookie to enable the feature to function properly.
22	in this case, we will notify you and keep you updated.
22	occasionally it may take us longer than a month if your request is particularly complex or you have made a number of requests.
21	we will retain and use your information as necessary to comply with our legal obligations, resolve disputes, and enforce our agreements.
20	how do we use your information?
20	alternatively, we may refuse to comply with your request in these circumstances.
20	how do we protect your information?
19	withdrawing your consent will not affect the lawfulness of any processing we conducted prior to your withdrawal, nor will it affect processing of your personal information conducted in reliance on lawful processing grounds other than consent.
19	we will respond to your request within a reasonable timeframe.
19	you will continue to receive generic ads.
19	unfortunately, the transmission of information via the internet is not completely secure.
19	for more information, please contact your local data protection authority.

please contact our u.s.-based third party dispute resolution provider (free of charge) at https://feedback-form.truste.com/watchdog/request.). A difficult part of this group is that sentences can only be attributed to this group if the privacy organisation is named; other sentences may also possibly be recommended by a privacy organization but because they aren't mentioned it cannot be proved in this research.

The last group is the group for which no logical reason can be identified why they are repeated, indicating that they possibly were copied. This is especially possible for the sentences with the highest frequencies. These contain some interesting sentences, such as *We consider this use to be proportionate and will not be prejudicial or detrimental to you*, a sentence which is specific when regarding that it was copied word for word 35 times. A common trend in this group is the topic of GDPR-related complex themes, such as access requests and lawfulness of processing. These sentences are likely copied as the websites were looking for specific formulations regarding a topic, and looked at other statements for ideas. Overall, this group indicates there is a substantial amount of copying between websites when creating privacy statements.

5.3.2 Statement Level Analysis

Apart from looking at specifically what sentences are copied, an additional analysis is also done to assess the state of similarities between statements. Of the 1510 statements analyzed, 287 had no shared sentences at all with any other statement. The average statement shared one or more sentences with 7 other statements, with a maximum of shared sentences with 177 other statements (by www.zazzle.com's privacy statement). This high number does not necessarily indicate copying, but could also indicate many companies copying their privacy statement.

The more interesting cases arise when taking the similarity metrics into account. For each statement, a similarity metric is measured for every other statement, which is a score based on the percentage of shared sentences ranging from 1, identical, to 0, no similarities (i.e. 1% shared sentences leads to a score of 0.01). Table 5.16 shows the percentage and frequency of the similarity metrics, giving an indication how many statements share sentences. An important note here is that similarities are measured for each statement and between each statement, leading to a total number of similarity scores of 1510 * 1509 = 1138540 scores. Table 5.17 shows the frequency of the maximum values of sentence similarity per statement. This table gives a more tangible idea of the amount of copying which is occurring between statements. To elaborate, statement in the range 0 have not a single shared sentence with any other statement. Statements in the group 0.2-0.4 have at least one other statement with which the similarity is between 0.2 and 0.4, all other statements have a lower similarity.

As can be seen in the table, the majority of statements (as expected) do not share any sentences. Small similarities often indicate a small number of identical sentences, which could be coincidence or a sentence 'inspired' by a different privacy statement. Very high values (0.8-1) often indicate statements which are both part of the same parent organisation, which chooses to use the same privacy statement for each company. The interesting cases are those with a similarity of 0.2 to 0.8. These represent a very small part of the full dataset, for which approximately half of the sentences are identical. For this group, statements are often very similar but the websites and companies behind these websites are not. A good example are the privacy policies from *9to5mac.com* [1], *AndroidPolice.com* [2], *cyberchimps.com* [3] and *themegrill.com* [4]. These privacy statements have many sentences with identical phrasings, but no logical reason why these statements would use the same formulation (such as a shared parent company or organisation).

However, a shortcoming of the information present is that it cannot clearly be distinguished if sentences are copied, and if so who copied these sentences in the first place. An interesting next step in this analysis would be to select the websites with a high likelihood of having copied other statements, and seeing what

[1] https://9to5mac.com/privacy/
[2] https://www.androidpolice.com/androidpolice-com-privacy-policy/
[3] https://cyberchimps.com/privacy-policy/
[4] https://themegrill.com/privacy-policy/

Table 5.16: Sentence Similarity Frequencies

Similarity	**Total**	**0**	**0-0.2**	**0.2-0.4**	**0.4-0.6**	**0.6-0.8**	**0.8-1**	**1**
Frequency	1138540	1126514	11539	171	97	73	146	0
Percentage	100,000%	98,944%	1,013%	0,015%	0,009%	0,006%	0,013%	0%

Table 5.17: Frequency of the maximum similarity value per statement

Range	**Total**	**0**	**0-0.2**	**0.2-0.4**	**0.4-0.6**	**0.6-0.8**	**0.8-1**
Frequency	1510	748	551	91	23	2	95
Percentage	100%	49,5%	36,5%	6,0%	1,5%	0,1%	6,3%

kind of group this is (i.e. what website category is highly present, what is their majestic rank, in what regions is more copying going on etc.). But the information present is not sufficient to identify such a group. When regarding the example in the previous paragraph, the information present is that *9to5mac.com* has similarities between 0.2 and 0.8 with 3 other websites. Although initially peculiar and interesting for analysis, *9to5mac.com* cannot be selected as a company which has copied other statements, as it may as well be true that they were copied.

5.4 Conclusions

In this chapter, the datasets achieved from the NLP-techniques in chapter 4 were analyzed in a number of ways. Firstly, analyses were done testing the differences in statements for each region and each website category. This revealed a large number of significant differences between regions and categories, of which a number lay in line with expectations and existing literature. The fact that these results aligned with literature gave an indication of the validity of the methods used. For instance, the regional analysis found that European privacy statements were mostly updated a year prior to the collection of the dataset, where shorter and generally more easy to read. This can be related to the GDPR which was adopted a year prior to the collection of the dataset. Furthermore, when checking the new dataset, European statements where more likely than American statements to mention terms related to the GDPR. North American statements had a significantly higher chance of mentioning third parties like Facebook, Google and Amazon. When analyzing the data per category, more findings can be done. News, Weather & Information websites have a relatively high chance of mentioning Double Click, Google's ad platform, which is in line with expectations. Websites in this category also mention terms related to the GDPR significantly less than other categories, indicating the possibility that these websites are lacking in their adoption of GDPR. Another interesting finding is that of Health & Fitness websites, which tend to collect relatively many different kinds of data, and have the highest chance of mentioning Google within their privacy statement.

After these initial analyses which served to validate the methods, a step further was taken and a number of companies where compared on a single-company level. The companies selected are the Big-5 tech companies, of which a number are known to have a high commitment to privacy, and others being subject to multiple privacy scandals. The data on these companies where compared to assess if the data could be linked in any way to the way companies treat data. If so, the expectation would be that companies who have a high commitment to privacy would have largely differing privacy statements than companies who are not committed to protecting the privacy of their users. However, upon inspection of all variables related to the privacy statements of the companies, no such relations were identified; this indicates that the variables do not extract information which can be directly related to a company's commitment to privacy.

In the following section, an attempt is made to see if the variables can be used to separate privacy statements based on their quality. This quality is assessed by using 3 metrics, which based on literature can indicate

if a privacy statement is well written and complete, or not. These metrics are informativeness (the extent to which a privacy statement covers all topics a privacy statement needs to cover), Flesch Reading-Ease score (the complexity of a text) and word count (the number of words, i.e. length, of the policy). This resulted in a number of best-case privacy statements (234 statements) and a number of worst-case statements (194 statements). The same tests as in the first two analyses were performed to assess if certain variables differed significantly between the good- and bad privacy statements. Firstly, the test resulted in a number of indicators which differ significantly between the groups. Best-case statements have a far higher likelihood of mentioning user-specific choices, specific audiences, how long data is kept, and if and how data is shared. An additional interesting finding was that the best-cases, according to the data, collected and shared significantly more data than the worst cases. Although possible, this result is unlikely to be true, as the websites were separated on informativeness and text usage, not on data usage. A far more likely conclusion is that the worst-case statements simply do not clearly mention what types of data they collect. This provides a meaningful addition, as it indicates that this method makes it possible to recognize inadequate privacy statements; a tool which could be of use for for instance national Data Protection Authorities (DPAs), to assess the state of the country's most important privacy statements.

A similarity analysis showed that there exist a lot of shared sentences between privacy statements. Some sentences of more than 15 words exist more than 20 times in the data. Furthermore, in 341 cases, a large number of sentences are shared with a high likelihood of being copied. Although it cannot be confirmed if a sentence is copied or not, a complex and long sentence which is used word-for-word multiple times indicates a high likelihood of copying. This copying indicates a number of companies do not prioritise their privacy statement and simply decide to copy it from other websites, indicating a lower commitment to privacy.

Chapter 6

Discussion

In this chapter, all steps taken in this research will be reviewed in the order that they were taken, and shortcomings encountered during these steps will be discussed. After each step, the implications of the shortcomings on the final results will be explained. This will provide better context for assessing to what degree the research goal has been met and the research questions have been answered, which will be discussed in the next chapter, chapter 7.

6.1 Methodology

First off a more general reflection is done on the research idea as a whole, concerning the discussion points which could already be mentioned based on the research as defined in the introduction and literature review. The goal of this research was to try and extract commitment to privacy from a company's privacy statement, based on the assumption that as privacy statements differ based on the effort a company put's into writing it, this effort must in some way be derivable and retracable to the original commitment to privacy. This section will take apart and analyze this assumption.

To start with, the privacy statement itself is a product of a socio-technical system, i.e. a system in which people and technology play an important combined role (Carayon, 2006), which is part of a far larger and more complex socio-technical problem, namely that of data-privacy. Furthermore, the privacy statement is a document containing natural language, which in itself is ambiguous and open to millions of variations (Bird et al., 2009). Although current regulations often have implications for the contents of privacy statements (see section 2.1.1), there is no uniformal format for privacy statements. Therefore, privacy statements are highly diverse documents, and by no means logical to extract information from. From these largely varying documents, an attempt is made to extract a company's commitment to privacy. Privacy commitment in itself is also a complex factor to measure, which is based on a combination of a company's moral values, resources, business model and data operations.

On the other hand, the methods used to extract this information (being Natural Language Processing algorithms created in Python) are of very precise nature; for binary variables, the algorithms use strict cut-off points to decide if a variable is true or not. This is to some degree true for each variable which is extracted using the NLP-techniques in this research. Furthermore, this is also true for the analyses performed. Tests are done to measure significant differences between groups of privacy statements, where a significant difference is defined by a strict cut-off point (alpha-value of 0.05).

Using NLP in this case also essentially reduces the amount of information which is present in the privacy statement to a select number of variables. Natural language contains endless amounts of syntactic and

semantic properties which can be extracted, of which only a select few are chosen to be extracted here. Although the variables are carefully selected based on which variables are expected to have the largest relevance for the research, a lot of information is still lost which could hold relevance for this research. The choice is therefore made to cover as much of the syntactic variables as possible for analysis, which can therefore function as an adequate first step. Syntactic analysis covers the essential direct elements of a text, while semantic analysis goes deeper into specific topics, such as sentiment analysis or build-up of grammar. Furthermore, this loss of information is inherent to a large-scale analysis of privacy statements as only a limited amount of information can automatically be analyzed.

This all combined entails that in this research, strictly scientific methods are applied to a set of diverse documents which are products of complex socio-technical systems. This has strong implications for the research. As can be expected, results are not to be seen as true measures or to be used as absolutes truths, but more as indications which hold value when used together. Eventual results are approximations of constructs, and by no means can be used to answer a clearly defined hypothesis for a single privacy statement. For instance, as each algorithm varies in accuracy, results from a single document hold less value than results from a large group of privacy statements combined. Furthermore, when interpreting the analyses, results which were slightly significant (alpha-values close to 0.05) are not relevant in the research, as the degree to which the results can vary is too large. Only very prominent trends with extreme significance bare relevance, and hold potential for further investigation.

Although initially this could be seen as slightly disappointing, it is inherent to the type of exploratory work which is performed here. Having a global idea of how companies are committed to privacy can be hugely useful for Data Protection Authorities and governments alike as it can be used to more efficiently apply regulations (which will also be further discussed in the conclusions, chapter 7). To the best of the researcher's knowledge, no significant work has been done towards extracting a company's commitment to privacy from a privacy statement. Because of this, some form of groundwork has to be laid in order to form a starting point for further research. As commitment to privacy is such a vague construct it cannot be easily extracted: current attempts have involved performing interviews with multiple companies, but these type of methods are lengthy, and are not scalable. This research forms the first attempt to extract this information on a large scale. This novelty combined with the limited scope of this research (as it is a masters thesis), is naturally paired with the drawbacks explained in the previous subsections. However, this research does contain multiple promising results for future work, which will be further discussed in section 7.3.

A second limitation which can be attributed to the methodology is related to the honesty and transparency of the company. In this research, for many variables honesty and transparency is assumed from the company in their privacy statement. Although this is assumed to be correct in the majority of the cases, in some cases it can simply be true that the company either deliberately or accidentally is not fully transparent in their privacy statement. Using this methods, certain types of data sharing can possibly not be mentioned, leading to incorrect values in the database. This is an issue which cannot easily be solved, and is accepted for this research. Possibly, regulators informing companies of incomplete privacy statements could increase the overall transparency as these companies are motivated to reassess their data practices and rewrite their statements. If that does not happen, this factor is seen as a fixed limitation of this type of research.

6.2 Data Collection

This section will firstly discuss all limitations encountered during the creation of the database of privacy statements, followed by the implications of these limitations for the results.

Limitations

Flattened Text Due to the style of the input used in the JavaScript code for Amazon MTurk, the text which workers copied was automatically flattened (i.e. all line breaks and formatting were removed). Despite multiple attempts at retrieving the segmentation of the text using Python, eventually no clear distinction could be made between titles and bodies of text. This lead to limitations in the research as the statement could not be divided into different subsections. Having kept the formatting could have further improved the extraction of data from the statements, by for instance only searching for specific information in the correct subsection. Furthermore, the formatting of a privacy statement could potentially give indications into what effort a company puts into making their privacy statement readable; a metric which could have been of importance for the research. This was also taken into account by Linden et al. (2018), who defended that presentation of a privacy statement is a direct indication towards the effort a company puts into readability. A problem recognized later in the research was that bodies of text like tables and lists were also flattened. In rare cases this would cause errors in metrics like the average sentence length, if the table was recognized to be one long sentence.

Majestic Million Another point of discussion is the use of the Majestic Million database. Although the database was filtered for the 2000 most visited websites, the database still contained many websites which were heavily outdated, or even offline. This lead to doubts on the workings of the database, and how recent the data used really is. What would seem to be a safer option, based on earlier research on the creation of privacy statements (Ramanath, Liu, Sadeh, & Smith, 2015), would be to use Alexa.com as a source of most popular websites. Although not free, this database is used for multiple forms of research, which do not seem to have an issue with outdated websites.

Worker Location A limitation which was recognized after the collection of the database was the impact of the location of the worker. As described in chapter 3, the database of privacy statements was collected by creating a task using Amazon Mturk. This task could be completed by workers worldwide, who would visit a website, enter the link of the privacy statement and copy the privacy statement itself. An effect which was not anticipated for was that the location of the worker would in some cases influence the website visited and the contents of the privacy statement. For instance, if a worker would click on the link *www.airbnb.com* from India, they would automatically be redirected to *www.airbnb.in*. This specific example was often encountered in the results, and clarifies the size of the Asia/Oceania group.

Website Categories Another shortcoming was identified during the category analysis, namely the largely skewed nature of the data. Although selected the top 2000 most popular websites were expected to have somewhat equally large representations for each website category, this was not the case. The data was largely skewed towards the category *'Technology & Computing'*, followed by an above average representation of the categories *'News/Weather/Information'* and *'Business'*. The reason for these skewed results could either be attributed to the tool used to extract the website categories, or the nature of the database. The effects of the tool can be seen when looking at the examples of *www.google.com, www.acer.com* and *www.appleinsider.com*. These three websites all belong to the category *'Technology & Computing'*, while in reality these could also for instance be divided into categories *'Search Engine'*, *'Retail'* and *'News'* respectively. This division is decided by the tool used, Webshrinker, which therefore has a significant impact on the results. Another possibility could also be the nature of the data which is analyzed. As the top 2000 most popular websites are analyzed, a lot of the companies behind these websites have their website as their main product (for instance websites like *www.google.com* and *www.appleinsider.com* are websites run by companies which have the websites as their main product or service). This would inherently mean that the company behind this website is a tech-company, as their main task is running websites.

Date last-updated collection The last shortcoming that had an impact on the research was the method through which the "last-update" date was collected. This date is the date which the statement was last

updated, which is often mentioned at the beginning or end of a privacy statement. Eventually this date was of great importance to the research as it defined the "new" database, which was primarily used for the analyses. The task of entering this date was given to the Amazon MTurk workers, who had to enter the url of the privacy statement, enter the date last-updated and the full text of the privacy statement. However, this date was often entered partially incorrectly or not entered at all. Because of this, far more statements did not have a date last updated than was in reality the case, leading to the 'new database' being smaller than necessary. Ultimately, writing an algorithm to retrieve this date would have most likely led to more accurate results, and therefore a more clean and useful database. Due to time limitations this could not be performed within the scope of this research.

Implications

Some of these shortcomings have implications for the final conclusions, others can be handled within the research itself. The fact that lay-out could not be extracted has serious implications for the results, as it reduces the number of results that can be achieved. Lay-out could have proven to be an interesting signal for privacy commitment, but this could not be analyzed in this research. Although this does not change any of the conclusions made in this research, it does reduce the level of detail to an unknown degree. For future research it would definitely be recommended to attempt to also extract lay-out, ideas for which will be presented in section 7.3.

The fact that a large number of websites seemed outdated was taken into account by also using a new dataset, which checked for the date a statement was last updated. The downside of using the Majestic database was that other databases could have provided a larger number of usable privacy statements than the database used for this research. Again, this does not impact any of the found results, although more privacy statement could potentially have enhanced the quality of results.

The worker location problem is a problem which is inherent to using Amazon MTurk, but only seldom lead to actual problems. To clarify, a number of possibilities arise when a privacy statement is visited from a different region for the same site. One option is that the same privacy statement is shown, which is most often the case. If this happens, no problem arises as the location of the worker does not influence the contents. A second option is that the privacy statement does change based on the region to apply to differing international regulations. This can potentially lead to problems, for instance when testing for GDPR variables; while the European version of the privacy statement may contain many GDPR-related terms, the Asian version may not, leading to the false implication that that website likely has not been edited for GDPR. This was in most cases solved with the regional analysis; when the privacy statement changed per region, the headquarters of the company would also change to the local headquarters, and the local region of the worker would therefore be identified. Therefore, performing the regional analysis made it possible to only test European statements for GDPR-variables, for instance. Based on manual assessment, a small number of statements kept their worldwide headquarters, leading the the statement-region being wrongly assigned. As mentioned, these situations were rare, and therefore accepted as a limitation. Other than some noise, this should not lead to large implications for the results.

The skewness of the website categories is a limitation which is accepted for this research, and for which not a lot can be done within the scope of this research. The skewness mainly leads to a loss in granularity for the categorical research, as many categories were underrepresented in the analysis and had to be left out. Performing the same analysis with all categories evenly divided would make it possible to more obtain more results based on how different website categories treat privacy. A next question which can then be asked is if the database used in this research is then representative for the average website as the categories are so skewed. This is highly probable, as the websites were chosen from a list of most popular websites, and not on category. As this is completely separate from category, it still represents a database of all websites. Also, as it contains the most visited websites, the database also contains the statements which a random internet user will most likely come in contact with, which is also the most relevant for this research. Therefore the skewness of categories is not expected to have any significant impact on the results, other than reducing the

granularity for the categorical analysis.

Lastly, the inaccuracy of the date last-updated variable mostly only implies less statements could be used in the eventual research than were possible. A large number of statements did not have dates entered leading them to automatically be set to the old database, while in reality these might have been suitable for the new database. In some rare cases the day and month would be switched (only possible if the day of the month would be the 12th or earlier, other values were fixed). However, as the year was always accurately retrieved this did not lead to large inaccuracies. For future work it would definitely be recommended to write this algorithm with hand, as that would lead to more accurate results and a larger usable database. As described earlier, a smaller database can lead to less retrieved results, and possibly could have further strengthened existing found results.

6.3 Natural Language Processing

During the extraction of the variables from the privacy statements a number of shortcomings were also identified. These are mentioned here, followed by their implications for the research.

Limitations

Data collection and -sharing purpose One of the goals of this research was to extract the purpose of collecting and sharing data. This was important as it implies the way the data is used, which in turn can function as an important signal for surveillance capitalism. Collecting a large number of types of data for marketing purposes could for instance form a signal for surveillance capitalists, as this is essentially the business model of a surveillance capitalist; collecting as much data as possible for profit. While extracting the variables, it turned out that extracting the purpose of collecting and sharing data was a lot more complex than the types of data being collected and shared, and that extracting this in an accurate manner was not feasible within the scope of this project. Because of this, identifying signals of surveillance capitalism was limited to extracting information regarding data operations and the other variables which were available for this research.

Techniques This limitation is already partially discussed in the first section of this chapter, section 6.1. Due to the scope of the research and the unlimited possibilities NLP research implies, a choice had to be made for what types of NLP-techniques to be used. Due to the exploratory nature of this research, the choice was made to perform mostly syntactic NLP-analysis and recreate variables from existing research. This fits within exploratory research as it provides an essential first step of analyzing the basic attributes of the data (syntactic analysis) and assessing what known variables can tell us. Based on the results of these variables next steps can be planned. Skipping these steps and immediately performing semantic analysis would be potentially using a difficult method to find a solution which can more simply be found.

NLP validation The method of validation is a quite limited form of validation; the chosen method is manual validation of ten random privacy statements for each variable. The outcome of the variable is then checked for accuracy by manually reading the privacy statement and seeing if it is in line with the result of the algorithm. This in itself is a very simple form of validation, as only ten privacy statements were checked. The amount of privacy statements (ten) is chosen as it is a reasonable number to use when applying manual validation due to the time needed for manual validation. Furthermore the manual assessments are fully performed by one person, which makes the assessments (and therefore validation) prone to bias; the same rules are applied to creating the algorithm and testing it, making it possible to miss certain aspects of what factor the algorithm is actually trying to measure. Lastly, only accuracy is measured in these tests, and

not more elaborate metrics like precision and recall, which give better insights into the performance of each algorithm. Although only measuring accuracy is deemed accurate for this research, further research could benefit from also using these additional measures to further enhance the accuracy of the algorithms and therefore quality of the results.

Inherent to using this method of validation is the impact of the chosen minimum accuracy level. For each variable extracted, the manual validation was tested to have a minimum accuracy level of 80%. This means that in worst cases, a variable can have approximately 20% incorrect input. A varying level of accuracy is not uncommon in NLP research, but as the method of validation was also limited, the true scale of incorrect values could potentially be higher. This is partly also one of the reasons that as explained in the introduction, results that can be derived from this research are notable trends, not individually comparable statistics.

Implications

The first two limitations impact the quantity of the results. If extracting purpose of data collection and -sharing was possible, the chance of finding a signal for identifying surveillance capitalists would have grown. Likewise, applying more different NLP-techniques can lead to more variables, therefore more possible analyses and more possible results. Both these limitations are related to the scope of this project, and leave possibilities for future work. In section 7.3, these possibilities will be presented.

The limitations regarding validation and accuracy have implications for how the results in this research can be used. These implications are also mentioned in the first section of this chapter and are also related to these limitations, namely regarding the way the results can be used. As the method of validation is not very secure, prone to bias and allows a certain level of errors, results are most suitable to be compared on a group basis. Comparing variables on an individual basis can be inaccurate as the chance that one or multiple of the tested variables are inaccurate is present. This effect is expected to smooth out when comparing variables on a large-scale basis.

6.4 Data Analysis

Two main limitations can be identified from performing the analyses, which are discussed here.

Limitations

Untrained analyses The first shortcoming regards the type of analyses which are performed. In this research, mostly methods are applied which are untrained, entailing that trends are sought within the data without any external influence. Although these types of analyses are a logical first step as they are relatively easy to perform, they are far from exhaustive. The database therefore still contains many possibilities for analyses which are not explored in this research: ideas for further analyses are discussed in section 7.3.

Clustering The second shortcoming regards the clustering analysis. While initially clustering was thought to be an interesting research step with a probability of obtaining results, this turned out to be less promising than expected. Firstly, a limitation which is also shortly highlighted in section 5.2.6, is that of the nature of the data. Clusters are most apparent in data which is mainly numerical, however the obtained dataset is mainly boolean. Furthermore, the full dataset was high dimensional, with 72 variables. Although the information extraction from the statements was maximized within the scope of this research, the resulting data was not suitable for clustering. In an attempt to improve this, a large number of boolean variables were

summed to create a number of ordinal variables. Although using ordinal and numerical variables is more suitable for finding clusters within data, these variables did not yield clusters.

Implications

The limitations encountered during the analyses have implications for the amount of results found, but not on the results themselves. If these limitations weren't encountered this research might have produced more useful results, but as this work is exploratory these kind of limitations can be expected. The first limitation is mostly a time and scope issue of this research, but gives great opportunities for future work. And second, although the results of the clustering were disappointing, the attempt at clustering was thorough, leading to the ability to conclude that the probability of clusters existing in this dataset are low, which in itself is also a conclusion.

6.5 Overview

The limitations discussed in this chapter can roughly be divided into two groups, which is shown in table 6.1. The two categories are defined limitations which impact the quality of the results, and limitations which impact the quantity of results. The groups can be interpreted as follows: limitations which impact the quality of results have an impact on the results of the analyses which are presented in this research. They impact the accuracy and therefore validity of the results, and therefore limit their predictive power. The other group, limitations which impact the quantity of results, are limitations which upon avoiding could have led this research to produce more different results due to a limitations of the amount of data available, or limitations to the amount of analyses possible.

Table 6.1: Discussion points Overview

Fixed Limitation	**Impacting quality of results**	**Impacting quantity of results**
NLP-techniques on socio-technical system	Worker Location	Untrained Analysis
Honesty and Transparency	Date last-updated	Semantic Analysis
	NLP-validation	Clustering
	Website Category	Flattened Text
	Majestic Million	

What can be taken from this table is that the limitations in the left column are limitations which can (to differing degrees) be improved for future research, in order to enhance accuracy of results. It is therefore important for future work to take these factors into account when performing research. The first limitation, mentioned in the first section of this chapter, forms an exception and is a limitation which is inherent to this type of research. The remaining limitations are possible improvements for this work, which were not pursued either due to the scope of this research, or due to differing results than expected. For this research, these limitations show that in the final results noise can be expected, and therefore results must be carefully interpreted. Predictive power comes from either combining a large amount of statements on one variable, or combining variables. The limitations stated in the right column are possibilities to further enhance the data and enrich the amount of results obtained from this data. In section 7.3, ideas of how future work can improve on this work and subvert these limitations are provided.

Chapter 7

Conclusions

From a privacy perspective, knowing how companies use data is becoming increasingly important in today's world. The way a company uses data which is collected from users defines if their users are subject to potentially large privacy risks, or if a user can hand over their data in good faith, knowing the company will keep it secure and to themselves. For most companies, no clear indication exists to which of these two groups they belong. The best indication that currently exists is a company's privacy statement, which gives some insight into how a company uses data. This research has attempted to use the information from privacy statements in order to clarify this problem, and provide a first step into discovering what can be learned about companies from their privacy statements.

This section provides an overview of the findings of this report. First, each subquestion is answered, based on work performed and described in this report. Next, the main research question is answered, based on the 6 subquestions and the research performed, followed by the contributions of this paper.

7.1 Evaluating the Research Subquestions

7.1.1 Findings

Question 1 *What do privacy statements contain, and on what aspects do they differ?*

A privacy statement is a body of text written by a party to inform the users of their services or products in which ways data is gathered, used, shared and managed. In general, a privacy statement's contents are defined partly by what the company deems necessary to notify the user of in terms of liabilities regarding data, but mostly by existing local regulations. Wilson et al. (2016) defined a taxonomy for privacy statement topics, which also serves as a requirement list for privacy statements. These ten topics are the most common contents of a privacy statement, and indicate what a privacy statement should contain.

Privacy statements differ on many aspects; this is apparent from existing literature and confirmed in this research using a new database. As a privacy statement is a written document with no set structure, the aspects on which privacy statements differ far outnumber the aspects on which privacy statements are alike. As mentioned above, a functional privacy statement informs the users in which ways their data is being gathered and used. The contents therefore rely on the way the company uses data, and the local regulations to which the statement must comply. The way the privacy statement is written and the structure of the statement relies on the company itself.

Question 2 *What variables can be extracted from privacy statements, using Natural Language Processing? What variables can be extracted within the scope of this research?*

Privacy statement analysis is not new; many research has been done into extracting information from privacy statements using natural language processing (NLP). As the ways in using NLP are abundant, the possibilities of information extraction are likely to be more than has been done up until now. To keep an overview and provide a tangible answer to this question, the results will be based on variables which have been extracted in previous research.

Of the research done in regards to automated analysis of privacy statements, many have taken concepts from earlier research, and through reiteration have added new concepts to the existing concepts. For instance, the coverage metrics defined by Wilson et al. (2016) are further expanded in the research by Harkous et al. (2018), and the concepts from this research are used to define new variables in research by Linden et al. (2018). The latter two represent elaborate work in privacy statement analysis, and extract a multitude of different, relevant variables from privacy statements. In this research, all concepts defined in these two works are therefore adopted as an overview of what has been extracted in the most recent literature. The variables are divided into 'Variable groups', and an overview of these groups is given in table 2.2. This table provides a tangible overview of what variables can be extracted from privacy statements.

The second part of this question entails discovering which of the found variables can be extracted within the scope of this research. To answer this question, firstly a new privacy statement corpus was created based on the top 2000 websites, filtered on top-level domain. After pre-processing and cleaning, this resulted in a database of 1510 unique privacy statements. Subsequently, a large number of NLP-techniques were applied to a subset of this dataset, to obtain the maximum amount of variables within the scope of this research. This resulted in 72 unique variables, of which 63 were extracted from the privacy statement using NLP. Applying these techniques to the full privacy statement corpus resulted in a large database of numerical variables attributed to the privacy statements, therefore resulting in measurable differences between privacy statements. The variables which could not be retrieved were sharing- and collection purpose of data; this information turned out to be too ambiguously and diversely stated in each statement to gain accurate results within the scope of this research. All other variables were either fully or partially extracted.

Question 3 *What information can be extracted from the variables and the privacy statement corpus? Are the results of the analyses in line with existing literature and expectations? Can information be extracted which signals privacy commitment, or surveillance capitalism?*

To explore what could be done with the database, a number of analyses were performed in chapter 5. First, an analyses was performed to assess if privacy statements differ per region. This yielded a large number of significant results, and a large number of these could be attributed to the adoption of the GDPR, therefore lying in line with expectations. A similar analysis was performed for website categories, analyzing if significant differences could be measured between the average values of websites categories. These were found, and a large number were also in line with existing literature or expectations. News websites had a higher chance of mentioning Double Click, an ad-service company owned by Google, and had a lower chance of mentioning GPDR-related topics. Both conclusions were supported by other research, indicating that the NLP-techniques were working as expected.

To find out what else could be extracted for the data, a multitude of other analyses were performed. Those that provided interesting results were further discussed in chapter 5. The first of these was a manual assessment of the big-5 tech companies (Facebook, Google, Amazon, Microsoft and Apple). As Apple is a company with a proven high commitment to privacy, and others (mainly Google and Facebook) have a lesser reputation for handling users data, the manual assessment could give a first indication if these companies, with clearly differing commitments to privacy, also clearly differ in the data. Against expectations, Apple's privacy statement was the hardest to read, most vague, and stated collecting and sharing the most different

types of data. As mentioned in the discussion, this conclusion needs to be taken with caution as stating the sharing of data is not the same as actually sharing data; an important limitation which is further discussed later in this chapter.

Another analysis which was done in an attempt to extract commitment to privacy is the extraction of privacy statement quality. Three variables were chosen which, based on existing literature, could indicate the quality of a privacy statement. These were informativeness (based on research by Wilson et al. (2016)), Flesch Reading Ease (a metric used to describe text complexity), and length of the statements. Applying these variables made it possible to select a number of best- and worst cases, and identify what variables formed the main identifiers of the groups.

Another possibility which is assessed in this research is that of the presence of clusters. If clusters would exist in the data, it would indicate that multiple statements had similar attributes which are significantly different from all other statements. This would in turn indicate that a number of companies either had strongly similar privacy statements, or companies had similar methods of handling data. To test for the presence of clusters, multiple methods were used on multiple versions of the data. Despite rigorous testing, all attempts did not result in valid clusters with any relevant meaning. Although this could also possibly be due to the format of the database, it is concluded that for this database it is highly unlikely that clusters exist. However, the possibility is not ruled out that upon extracting different variables of privacy statements, relevant clusters might exist.

To further gain insights how companies treat privacy, an attempt is done to see if statements are copied by performing a similarity analysis. For this analysis, the number of shared sentences is measured for each statement. This resulted in an overview of the most frequently used sentences and which statements shared a large number of sentences with other statements. The results indicated a high likelihood of copying: some (long) sentences were used over 20 times in the database, and some statements shared more than half of the sentences with other privacy statements without sharing any other connection with the company that wrote the other privacy statement.

Question 4 *Do the analyses lead to results with implications for governments and regulators? What can be learned from these implications?*

To answer the final question, a final reflection is done on the results achieved in this research, and the answers to the previous research questions.

7.1.2 Implications of the Findings

Now the first three research subquestions have been tackled, the final research subquestion, regarding the implications of the findings, can be answered. To further guide this process, the conceptual framework defined in section 2.4 is used to assess what part of the research has succeeded, what parts need further research, and what implications can already be defined.

Reflecting on the conceptual framework

As can be seen in figure 7.1, the research was roughly divided into three steps which are shown as the dotted arrows in the right part of the conceptual framework. The research goal was to firstly attempt to extract a company's attitude towards privacy using the variables extracted from the privacy statement. Next, if that was possible, assessments needed to be made towards seeing if the data operations could be extracted, and if so possibly also the business model; this step was important as extracting the business model and data operations could potentially indicate surveillance capitalism. Lastly, based on these findings, the possibility of improving regulations would be assessed; this last step will be done in this section.

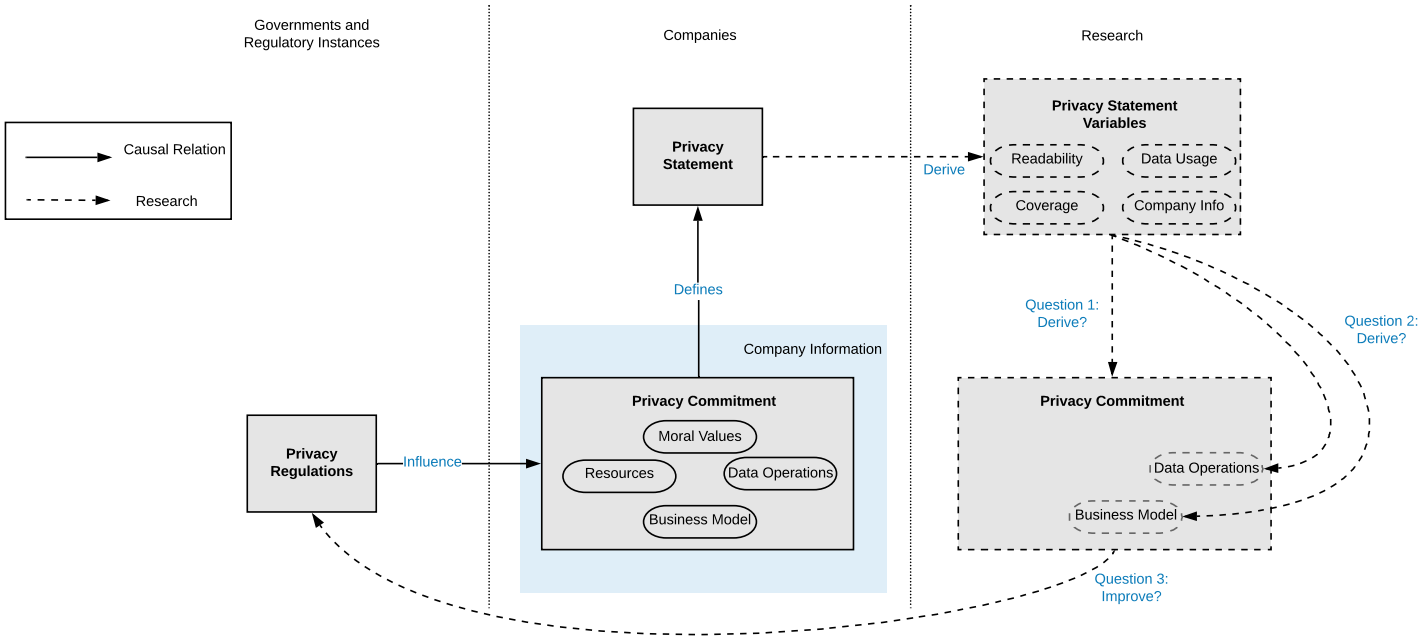


Figure 7.1: Conceptual Framework

Commitment to Privacy Two attempts were made towards extracting a company's commitment to privacy, namely the statement quality analysis and the similarity analysis. These analyses succeeded in uncovering if a privacy statement was complete and well written or not, and to what degree a privacy statement had identical sentences, possibly indicating the statement was copied. These factors do to a certain degree relate to commitment to privacy; a company which copies their privacy statement or writes an incomplete privacy statement has not put enough commitment into writing the document which serves to protect the users privacy. These measures are not very precise and exist mostly by comparing to the average privacy statement in the database, but still prove promising to assess whether a privacy statement is complete and easily readable, or needs improvements.

These two tools can function as an indication to the amount of effort put into a privacy statement. This is therefore an important finding for regulators (Data Protection Authorities); if regulators wish to assess to what degree a group of companies have put effort into their privacy statement, this tool can give an indication of that effort. Also, as elaborated in section 2.1.4, the privacy statement forms the first step in a company's strategy towards privacy, consequently providing an indication to commitment to privacy. As it is the task of DPA's to supervise privacy law (European Commission, 2017), this tool can hence be used for DPA's to more efficiently fulfil that task, as finding companies which do not adhere to regulations is made easier. An important note to be taken with this measure is that it is up until now fully based on literature, and would require some form of external validation to test its accuracy, which is a possibility for future work.

Furthermore, this research is by no means exhaustive. These initial analyses show promising results towards extracting commitment to privacy and leave additional possibilities to further explore this field. More in depth analysis of the text could for instance extract contractual issues, unjust use of personal data, third-party tracking and other aspects which possibly infringe privacy. This would give a more detailed idea of how companies treat privacy, in stead of simply the effort put into the privacy statement. If successful, this information can further guide regulators in effectively defending consumers privacy, as this research can more clearly show in what ways it is infringed.

Surveillance Capitalism
However, the research goal of this research was not to create a tool which can be used to police companies which have inadequate privacy statements. The main problem arose from surveillance capitalism, where companies would collect massive amounts of data for their business model, infringing consumers privacy. That is where the second step comes in in the conceptual framework of attempting to extract features like data operations, and with that possibly the type of business model the company has. A first limitation recognized here is that with the variables available, extracting business model wasn't possible as this is linked to the purpose of collecting and sharing data. Within this research, only types of data collected and shared were extracted, while extracting purpose was not feasible within the scope of this project.

Therefore, assessment could only be done based on the data operations and all other variables. The first analysis attempted was to compare companies that are well-known surveillance capitalists, with similar tech companies who show a higher commitment to privacy based on the variables. Results were contrary to expectations, with the privacy statements of the surveillance capitalists being easier to read, clear and informative. Clustering analysis was performed with the goal of finding companies with similar data practices from their privacy statements, but due to multiple reasons explained in section 5.2.6 and chapter 6, this did not lead to useful results.

These first analyses showed that the variables themselves do not clearly indicate which company is a surveillance capitalist and which is not. This is an essential first step in this process: if it did lead to results, signals for surveillance capitalism could easily be identified and used by regulators. As the analysis did not lead to results it means that further analysis is needed to find out if surveillance capitalism can be identified from privacy statements. The method and variables used in this research were not adequate; therefore, further elaborating on the method (performing trained analysis, sentiment analysis) or the variables (extracting purpose of collecting and sharing data or other aspects of the text) could provide results. This will be further discussed later in this chapter, in section 7.3.

Further Implications
Because the second question could not be answered within the scope of this research, the possibilities to answer the third question are also limited. If the research would have produced a tool or identified a signal which can identify surveillance capitalists from their privacy statements, it would have become possible to gain a more tangible understanding of the amount of surveillance capitalists which operate today and how they work. This, in turn, could lead to insights in how to regulate these companies. But even completing this step would likely require additional research and would not have been feasible to complete within the scope of this research.

The most important implication which can therefore be derived from the research at this point is that commitment to privacy can be extracted from privacy statements. Currently this can only be done high-level, with strong indications of which statements are below average in quality, or have likely copied other statements. Further research can improve this metric to be able to measure more granular levels of commitment to privacy. Using similar methods and variables, signals hinting towards surveillance capitalism cannot be identified, but do possibly exist in the statements. Because of this, no implications can be made for regulators and governments yet on this subject; this requires future research, which will be discussed in section 7.3.

The bigger picture A last question which then can be asked is where this research fits within the status quo. As of writing this paper, the GDPR is now in effect for over a year and has a significant impact on the way companies handle privacy, which is also discussed in chapter 2. The largest challenge within the global issue of data privacy is still the issue which surveillance capitalism underlines, namely that companies have found methods which comply with all current regulations, to collect and use large quantities of data for their business model; this has widely been seen as largely infringing on privacy, as it requires and maintains a large knowledge asymmetry between company and user.

This paper has made conclusions in both areas, of which the most significant was towards improving commitment to privacy. But as this paragraph already suggests, the GDPR has already had a large influence in improving commitment to privacy, also confirmed in (Libert, 2018). Of course, having a tool which enables effectively finding regulatory laggards easier can be of use for privacy regulators, but within these two topics more pressing issues are to be solved, which this research can still prove to be useful for.

Within commitment to privacy, perhaps the issue is not the companies that are lagging in terms of their privacy statements, but more those which have thought there privacy statements through in a very detailed manner to benefit themselves. Privacy statements that contain all necessary elements, but force users to for instance sign away their rights, share specific data or to state certain aspects of the statement in such ambiguous language that it deliberately confuses the user. This type of behavior is also possible to extract from privacy statements, and extensions of this work can be used for that purpose. On the topic of recognizing surveillance capitalism, this work has made first steps but has not succeeded. Future work can dive more into extracting purpose of data collection and -sharing, and for instance look at using a trained, annotated dataset to further test the variables to recognize surveillance capitalists. This work therefore functions mostly as a stepping stone forward, towards solving the more pressing issues within these topics.

7.1.3 Answering the main research question

Now all sub-questions are answered, the main research question can be revisited:

Can a company's privacy commitment be derived from their website's privacy statement using Natural Language Processing? If so, can other underlying attributes of privacy commitment be derived which could indicate surveillance capitalism? What implications does this have for privacy regulators and governments?

This research question is now answered in the previous subsection, which is summarized below:

- **A company's privacy commitment can be derived from the privacy statement, at least at a high-level.** This research succeeded in finding a measure which can indicate commitment to privacy based on literature, and has potential to be further expanded to derive more low-level aspects of commitment to privacy.
- **With the variables present and the methods used in this research, it is not possible to derive attributes from privacy statements which indicate surveillance capitalism.** However, analysis on this part of the research question was limited and many possibilities for future work exist.
- **If interested, regulators can encourage additional research towards extracting commitment to privacy.** If regulators would be interested in extracting commitment in order to improve the way websites treat consumers data, these first results are promising. Additional work would be needed to turn this work into a validated, usable product.

7.1.4 Recommendations

The nature of this research was exploratory, with the goal of discovering the limits of what can be learnt using a new technique, namely uncovering information about the company from privacy statements. The research functions as a first try-out to see what methods hold promise, and what areas future research can further look at. Because of this, inherently not many recommendations for policy-makers can be made based on this research alone. However, the broader societal relevance of this thesis still stands, even though policy cannot be shaped from this work yet. The broader problem of finding working policy to regulate surveillance capitalism is still relevant and holds possibilities for large societal contributions based on policy recommendations; this work is a first step in working towards this solution.

One recommendation can be made which has been mentioned throughout this section, which is aimed at the regulators of privacy and not necessarily policy-makers is the following: **privacy regulators should encourage research towards using privacy statements as indications for commitment to privacy.** As it is the task of privacy regulators to supervise the compliance with privacy law, a tool which can easily identify companies which are likely not to comply can be of large use, as it makes regulating these companies more efficient. To be fully employable a new tool has to be created, which can be based on the techniques found in this research. Section 7.3 will further elaborate on this.

7.2 Contributions

This section highlights the contributions made by this research, which are divided into scientific and societal contributions.

7.2.1 Scientific Contributions

- This research forms a first attempt at extracting commitment to privacy from privacy statements. A connection is made in the literature review of how commitment to privacy relates to the privacy statement, and a measure is created to extract this commitment to privacy. The first results are promising, further validation can prove or disprove the effectiveness of this method.
- This research also has taken the first steps in attempting to extract aspects of surveillance capitalism from a privacy statement. In this process, many methods of analyses have been attempted in order to gain results, of which many did not yield results. Knowing which methods do not yield results shows from where future work can further progress.
- A new privacy statement corpus of 1510 statements, dating from after GDPR and before ePR is created and made available for further research. This data can be used for further research into the state of privacy statements, or other usages of privacy statements. The dataset is extracted in a unique position, as it provides the possibility of measuring the effect of the GDPR alone. It is expected that the ePR will also have significant effect on privacy statements, which can only be accurately measured if a database exists dating in between the two regulations. This database will be made available for all TU-Delft students, and also posted online, making it possible for other researchers to also perform separate analysis on the statements. Additionally, it is the only dataset of this size, post-GDPR, which is publicly available.
- This research has also lead to a number of methodological contributions. It has shown the level of accuracy which can be achieved when using relatively simple NLP-techniques on a database of privacy statements. Also, multiple methods have been evaluating to extract aspects of privacy statements, such as where the headquarters of a company lie, if the users are updated via e-mail when the privacy statement changes, the vagueness of a text and more. These metrics are all explained in chapter 4, and shown in appendix F.

7.2.2 Societal Contributions

This research has contributed to assessing more efficient methods of regulating companies which do not comply with privacy laws. Upon further testing, this method could be built into a useful tool for regulators to be able to work more efficiently in assessing commitment to privacy of companies, and making sure companies comply with the most recent privacy laws. In turn, this can contribute to the defence of privacy of citizens. Apart from this contribution, this work has a strong focus on being a stepping stone for future

research. As of such, the focus is mostly on providing scientific contributions and opportunities for future research, which in turn can provide possibilities for societal contributions and policy recommendations.

7.3 Future Work

This research has made first steps towards extracting information about companies from privacy statements. A database of 1508 privacy statements has been collected, and a large number of metrics defined in other research have been extracted from the text. A number of analyses have also been performed to assess which methods hold potential and which don't. All these steps are useful for future work, as they need not be repeated. A large part of this thesis has been collecting the dataset and extracting the variables, which will not need to be repeated in future work.

A number of different types of future work are defined, these range from improvements to the method used in this research to completely new ideas.

Improvements of the current method

- **Extracted the date the privacy statement was last-updated using code**
  An issue which had a negative influence on the quality of the dataset was the fact that the workers had entered the last-updated dates of the privacy policies. Although this was initially done to save time, it would have in hindsight been more effective to extract the dates from the privacy statements using natural language processing. Using similar methods as in this thesis would also have succeeded in extracting the last-updated date, likely with a higher accuracy. This would also have lead to a far larger new dataset, making it possible to analyze the data with increased accuracy.
- **Extracting privacy statements from urls using Python**
  Another possible improvement of the dataset can be achieved by extracting privacy statements using code. As the database contains links to the privacy statements of a large number of privacy statements, it could be possible to extract the html-code of these statements using Python. This would solve a number of issues: First, as all statements would be extracted from one location, the problem of statements changing due to the location of the worker would be solved, as all statements would be extracted from one location. Second, html-code also contains the layout of the text, which would make it possible to also build analyses based on how text is structured (an indicated shortcoming of this research). Lastly, it would make it possible to enlarge the database, as a number of websites without valid privacy statements did have valid privacy statement urls.
- **Level-out statements per region and category**
  A factor that influenced the analysis was that there was an uneven distribution of websites per site category and region. Although corrected in this research, it led to statements either having to be grouped together or dropped out, which is a loss of information. An approach which could have been taken was to select a far larger selection of the Majestic Million (or other website database), primarily analyze their region or site category, and then create the database based on an evenly divided group. This would give a clearer picture of how each group influences the variables, and lead to data not having to be left out in analyses.

Extensions of the data

- **Extract layout**
  The fact that the layout eventually could not be extracted from the dataset proved to be a potential shortcoming for the analysis, as layout has a significant impact on the clarity of a privacy statement. Extracting layout could therefore be a possible additional metric to improve the statement quality

analysis. Extracting it can for instance be done by using the method described above (scraping statements from the link), by looking into additional methods of copying statements for workers, or by looking into an input type supported by Mechanical Turk which keeps the layout intact.

- **Collecting purpose of data collection and -sharing**
  An important limitation of this research was the fact that eventually only types of data were extracted from the data, and not the purpose of collection and sharing. This is a difficult variable to extract as it is mentioned in numerous ways, but combined with annotation this could for instance be made possible. As the purpose of the collecting and sharing of data could indicate the business model of the company, this variable could be a signal for surveillance capitalism.
- **Apply further methods of semantic NLP analysis** This research has mostly been limited to syntactic NLP analysis, extracting features from the text which are (mostly) directly observable. Using semantic NLP makes it possible to extract more nuanced information. This could make it possible to identify companies that have unfair privacy statements towards their users, check if statements contain mostly "law-related" terms or clearly understandable language, or if privacy statements for instance use language which implies they are hiding information. The possibilities are countless, and combined with annotation could lead to interesting results.

Additional uses of current data

- **Data network analysis**
  As mentioned throughout this research, the privacy statements contain a large number of third parties. This research simply extracted the presence of these third parties, but additional analysis could lead to interesting insights into the way data flows. For instance, additional analysis can first be done towards how these third parties are mentioned. If it is possible to extract if data is being exchanged with the company, this information can be used as a branch in a network. Analyzing a multitude of statements would then make it possible to create a full network of data flows between companies, leading to insights in how these companies use data.
- **Presence of third-parties in statements**
  An interesting finding done in one of the analyses (section 5.2.3) was the high presence of certain third parties in certain categories of websites (for instance, Health & Fitness websites had the largest presence of Google in their privacy statements). Research could be done into what the role is of these third parties, what kind of data is being shared, and what this implies for the company collecting the data.
- **Creating privacy nutrition lables with NLP**
  Using relatively simple NLP-techniques it was possible to extract a large number of variables from privacy statements. A lot of these variables coincided with the elements of the privacy nutrition label, an idea defined in 2009 (Kelley, Bresee, Cranor, & Reeder, 2009) to let companies fill in an equivalent of a nutrition label, but focused on what data is collected and shared, and why. Ultimately the first attempt of this work did not succeed, as it required the website themselves to fill in the nutrition label, many of which did not do. However, using NLP it may be possible to fully automatically fill in the nutrition label using natural language processing. This would make information from privacy statements a lot easier to understand for users, and provides the possibility of reinvoke the discussion on the privacy nutrition label.
- **Supervised Machine Learning**
  This research has focused on unsupervised methods of analysis, where purely the data itself is analyzed to find patterns. As this did not yield results, a next logical step is to attempt trained analysis, whereby an annotated subset of the privacy statements is used. There are multiple possibilities of annotation within this dataset, two ideas are provided here:

- The first and most logical step would be to further expand on the surveillance capitalism analysis provided in this research. This would mean attempting to retrieve a list of companies or privacy statements which are known to be surveillance capitalists, together with a large number which are known not to be surveillance capitalist. This, in combination with the variables extracted in this research, would make it possible to clearly identify if, and if so which variables are signals towards surveillance capitalists. However, as surveillance capitalists are currently not well known, this annotation is no straightforward task.
- Another idea would be to have a number of privacy professionals annotate the statements. In this case, the statements could be annotated by how clear they are, how appealing their layout is, if their terms are fair etc.. This would provide a quailty measure from a professional viewpoint, which could be used to validate the statement quality metric provided in this research, or for instance to create a whole new quality metric based on other variables.
- Lastly, it could be possible to have the statements annotated by a random selection of people, to get an indication of what makes a privacy statement clear, readable or fair to a normal user. This could make it possible to quantify what makes a privacy statement 'good', or for instance see if there are parts of a privacy statement which are overlooked by the most people.

Of course, further methods of annotation are possible, and can for instance be inspired by existing research on annotation of privacy statements.

References

Acquisti, A., Taylor, C. R., & Wagman, L. (2015). The Economics of Privacy. *Ssrn*, 1–58. Retrieved from https://www.law.berkeley.edu/wp-content/uploads/2015/11/The-Economics-of-Privacy.pdf doi: 10.2139/ssrn.2580411

Altman, I. (1975). *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding.* Cole Publishing Company.

Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2009). Does Anyone Read the Fine Print? Consumer Attention to Standard Form Contracts. *Ssrn*. Retrieved from https://lsr.nellco.org/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=119 doi: 10.2139/ssrn.1443256

Beasley, T. M., & Schumacher, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post hoc and planned comparison procedures. *J. Exp. Educ.*. doi: 10.1080/00220973.1995.9943797

Bhatia, J., Breaux, T. D., Reidenberg, J. R., & Norton, T. B. (2016). A Theory of Vagueness and Privacy Risk Perception. In *Proc. - 2016 ieee 24th int. requir. eng. conf. re 2016.* doi: 10.1109/RE.2016.20

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*. doi: 10.1016/j.datak.2006.01.013

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.*

Brakenhoff, L. (2017). *Extracting Components from Privacy Statements with Text Mining* (Unpublished doctoral dissertation). Universiteit Utrecht.

Brandeis, S. D. W. L. D. (1890). The right to privacy. *Harv. Law Rev.*, *4*(5), 193–220. Retrieved from http://links.jstor.org/sici?sici=0017-811X%2818901215%294%3A5%3C193%3ATRTP%3E2.0.CO%3B2-C%0Ahttp://

Carayon, P. (2006). Human factors of complex sociotechnical systems. *Appl. Ergon.*. doi: 10.1016/j.apergo.2006.04.011

Cate, F. (2010). The Limits of Notice and Choice. (April), 59–62. Retrieved from http://www.lawtech.hk/pni/wp-content/uploads/2015/04/Fred-H-Cate.pdf

Catterall, M. (2000). *Research Methods for Business Students.* doi: 10.1108/qmr.2000.3.4.215.2

Costante, E., Sun, Y., Petković, M., & den Hartog, J. (2012). A machine learning solution to assess privacy policy completeness.. doi: 10.1145/2381966.2381979

Council of Europe. (1950). *Convention for the Protection of Human Rights and Fundamental Freedoms* (Tech. Rep.). European Union.

Cranor, L. (2012). Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. Telecommun. High Technol. ...*, 273–307. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jtelhtel10&section=22%5Cnh

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elem. English*. doi: 10.1111/j.1467-9345.1968.tb00749.x

Dolata, U. (2017). *Apple, Amazon, Google, Facebook, Microsoft: Market concentration - competition - innovation strategies.*

Dunn, O. J. (1961). Multiple Comparisons among Means. *J. Am. Stat. Assoc.*. doi: 10.1080/01621459.1961.10482090

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proc. 2nd int. conf. knowl. discov. data min.*

Esteve, A. (2017). The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. *Int. Data Priv. Law*, *7*(1), 36–47.

European Commission. (2017). EC Proposal ePrivacy regulation. Retrieved from `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52017PC0010&from=EN`

European Parliament. (2016). Regulation (EU) 2016/679 (GDPR). *Off. J. Eur. Union*, 1–88. Retrieved from `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0132.01.ENG&toc=OJ:L:2` doi: L:2016:119:TOC

European Union. (2016). Regulation 2016/679. *Off. J. Eur. Communities*, *2014*(March 2014), 1–88. Retrieved from `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679` doi: http://eur-lex.europa.eu/pri/en/oj/dat/2003/$l_2$85/$l_2$8520031101*en*00330037.*pdf*

Falessi, D., Cantone, G., & Canfora, G. (2013). Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. *IEEE Trans. Softw. Eng.*. doi: 10.1109/TSE.2011.122

Fellbaum, C. (2005). WordNet and wordnets. In *Encycl. lang. linguist.*

Fitzgerald, J. L., & Watkins, M. W. (2006). Parents' rights in special education: The readability of procedural safeguards. *Except. Child.*. doi: 10.1177/001440290607200407

Flesch, R. (1949). The Art of Readable Writing. *Stanford Law Rev.*. doi: 10.2307/1225957

Garriga, E., & Melé, D. (2013). Corporate social responsibility theories: Mapping the territory. In *Cit. class. from j. bus. ethics celebr. first thirty years publ.* doi: 10.1007/978-94-007-4126-3_4

Gower, J. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*. doi: 10.2307/2528823

Graham, J. (2018, apr). *Is Apple really better about privacy?*

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Clustering validity checking methods: part II. *ACM SIGMOD Rec.*. doi: 10.1145/601858.601862

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. Retrieved from `http://arxiv.org/abs/1802.02561`

Hernandez, G., Arias, O., Buentello, D., & Jin, Y. (2014). Smart Nest Thermostat : A Smart Spy in Your Home. *Black Hat USA*, 1–8. Retrieved from `https://www.blackhat.com/docs/us-14/materials/us-14-Jin-Smart-Nest-Thermostat-A-Smart-Spy-In-Your-H`

Hintze, M. (2017). In Defense of the Long Privacy Statement. *Maryl. Law Rev. (Baltimore, Md.)*, *76*(4), 1044. Retrieved from `https://pdfs.semanticscholar.org/7d26/d6cac2b19d023a1e3e9cb62d5fb7a64f6306.pdf`

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.*. doi: 10.1023/A:1009769707641

Human Rights Council. (2016). The promotion, protection and enjoyment of human rights on the Internet. *United Nations Gen. Assem.*, *10802*(June), A/HRC/32/L.20. Retrieved from `https://www.article19.org/data/files/Internet_Statement_Adopted.pdf` doi: 10.1093/oxfordhb/9780199560103.003.0005

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer (Long. Beach. Calif).*. doi: 10.1109/MC.2018.3191268

Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009). A "nutrition label" for privacy.. doi: 10.1145/1572532.1572538

Kincaid, J. P., Aagard, J. A., O'Hara, J. W., & Cottrell, L. K. (1981). Computer readability editing system. *IEEE Trans. Prof. Commun.*. doi: 10.1109/TPC.1981.6447821

Kincaid, J. P., Fishburne, J., Robert P., R., Richard L., C., & Brad S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (Tech. Rep.). doi: 10.21236/ADA006655

Kiyokawa, H. (1996). CHALL, J. S. and DALE, E. (1995) Readability Revisited : The New Dale-Chall Readability Formula., Brookline Books. *Japanese J. Educ. media Res.*. doi: 10.24458/jaems.3.1$_5$9

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.*. doi: 10.1080/01621459.1952.10483441

Libert, T. (2018). An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies.. doi: 10.1145/3178876.3186087

Libert, T., & Nielsen, R. K. (2018). Third-Party Web Content on EU News Sites: Potential Challenges and Paths to Privacy Improvement. (May), 1–11. Retrieved from `https://timlibert.me/pdf/Libert_Nielsen-2018-Third_Party_Content_EU_News_GDPR.pdf`

Linden, T., Harkous, H., & Fawaz, K. (2018). The Privacy Policy Landscape After the GDPR. Retrieved from `http://arxiv.org/abs/1809.08396`

Liu, F., Ramanath, R., Sadeh, N., & Smith, N. (2014). A Step Towards Usable Privacy Policy: Unsupervised Alignment of Privacy Statements. *Proc. 25th Int. Conf. Comput. Linguist. (COLING 2014)*.

McDonald, A. M., & Cranor, L. F. (2008). The Cost of Reading Privacy Policies. *Inf. Syst. A J. Law Policy Inf. Soc.*, *4*(3), 540–565. Retrieved from `https://kb.osu.edu/bitstream/handle/1811/72839/ISJLP_V4N3_543.pdf?sequence=1&isAllowed=y` doi: 10.1016/B978-1-59749-615-5.00013-X

McLaughlin, G. (1969). SMOG grading: A new readability formula. *J. Read.*, *12*(8), 639–646.

Miller, Jr., R. G. (1997). *Beyond ANOVA*. doi: 10.1201/b15236

Mysore Sathyendra, K., Schaub, F., Wilson, S., & Sadeh, N. (2016). Automatic Extraction of Opt-Out Choices from Privacy Policies. In *Aaai fall symp. priv. lang. technol.*

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proc. - IEEE Symp. Secur. Priv.*, 111–125. Retrieved from `https://www.cs.utexas.edu/ shmat/shmat_oak08netflix.pdf` doi: 10.1109/SP.2008.33

Nissenbaum, H. (2012). A contextual approach to privacy online. In *Digit. enlight. yearb. 2012.* doi: 10.3233/978-1-61499-057-4-219

Polanyi, K. (1978). *The great transformation: Politische und ökonomische Ursprünge von Gesellschaften und Wirtschaftssystemen*. doi: 10.2307/2144137

Pollach, I. (2011). Online privacy as a corporate social responsibility: An empirical study. *Bus. Ethics*. doi: 10.1111/j.1467-8608.2010.01611.x

Provost, F., & Fawcett, T. (2013). *Data Science for Business*.

Ramanath, R., Liu, F., Sadeh, N., & Smith, N. A. (2015). Unsupervised Alignment of Privacy Policies using Hidden Markov Models.. doi: 10.3115/v1/p14-2099

Reidenberg, J. R., D, S., & Callen, A. J. (2014). *Privacy Harms and the Effectiveness of the Notice and Choice Framework* (Tech. Rep.). Retrieved from `http://www.ftc.gov/sites/default/files/documents/reports/federal-`

Rosch, J. T. (2011). *DO NOT TRACK: PRIVACY IN AN INTERNET AGE* (Tech. Rep.). FTC.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*. doi: 10.1016/0377-0427(87)90125-7

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., Mcdonald, A. M., Reidenberg, J. R., ... Wilson, S. (2013). The Usable Privacy Policy Project : Combining Crowdsourcing , Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. *Tech. Rep. C.*.

Schultz, J. K., Yaqub, S., & Oresland, T. (2016). *Privacy policies of android diabetes apps and sharing of health information: In reply.* doi: 10.1001/jama.2015.17873

Schwartz, P. M., & Solove, D. (2009). Notice & Choice. *Second NPLAN/BMSG Meet. Digit. Media Mark. to Child.*. Retrieved from `http://digitalads.org/documents/Schwartz_Solove_Notice_Choice_NPLAN_BMSG_memo.pdf`

Shields, P. M. (2003). A Pragmatic Teaching Philosophy. *J. Public Aff. Educ.*. doi: 10.1080/15236803.2003.12023567

Tene, O., & Polonetsky, J. (2012). Privacy in the Age of Big Data: A Time for Big Decisions. *Comput. Law Secur. Rev.*. doi: 10.1145/1125170.1125230

W3C. (2019). *Tracking Preference Expression (DNT)* (Tech. Rep.). Author.

Ward, J. S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions. Retrieved from `http://arxiv.org/abs/1309.5821`

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Giovanni Leon, P., ... Sadeh, N. (2016). The Creation and Analysis of a Website Privacy Policy Corpus.. doi: 10.18653/v1/p16-1126

Zamanian, M., & Heydari, P. (2012). Readability of Texts: State of the Art. *Theory Pract. Lang. Stud.*, *2*(1), 43–53. Retrieved from `https://pdfs.semanticscholar.org/3adf/9a2d0d9579e3f688dd660c28a657fa55cead.pdf` doi: 10.4304/tpls.2.1.43-53

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *J. Inf. Technol.*, *30*(1), 75–89. Retrieved from `https://link.springer.com/content/pdf/10.1057%2Fjit.2015.5.pdf` doi: 10.1057/jit.2015.5

Zuboff, S. (2019). *The Age of Surveillance Capitalism* (Vol. 1) (No. 1). Profile Books.

Appendix A

Clustering

This Appendix discusses the methods used to assess for clusters within the dataset. Clusters are groups of data-points which have a significantly shorter distance to each other than to data-points outside of that groups (Provost & Fawcett, 2013). In reality, a cluster in the dataset would indicate that there are one or multiple groups of privacy statements which are similar in a number of aspects (these aspects are based on the variables extracted), and mainly significantly more similar to each other than to other privacy statements in the dataset. This could for instance indicate that a number of companies use the same method of writing a privacy statement, or even more interesting, companies have a similar way of using data. This appendix therefore assesses to what degree these clusters exist within the data.

A.1 Methods

Before applying clustering analysis it is important to understand the nature of the dataset. Depending on the dataset, different methods of clustering can be used. The dataset extracted from the privacy statements is a dataset with a total of 72 variables, all of mixed nature (i.e. consist of numerical, ordinal, categorical and boolean variables). Also important to note is that some variables are related to each other. Data collection, for instance, is an ordinal variable which consists of 7 boolean variables. Clustering variables which are dependent on each other can yield false clusters (Provost & Fawcett, 2013), thus it is important to take the correct variables into account when applying the analysis.

A.1.1 Clustering Algorithms

Because the data is mixed, the data is divided into subsets, one containing all numeric variables and the ordinal variables (which are composed from 21 boolean variables), and one full dataset, containing all numeric, boolean and categorical variables without the ordinal variables (as these might cause dependency issues as described above). Next, 3 clustering algorithms are chosen; one which is specifically made for numerical and ordinal data, one which is specifically made for mixed data, and one which can handle both. Specifically these methods are chosen as they provide the possibility to assess for different types of clusters in a dataset, enhancing the chance of finding results.

The first method is K-means (Provost & Fawcett, 2013). K-means is one of the earliest and most straightforward clustering methods. K-means works by randomly selecting the locations of a predefined number of centroids, and assessing the distance between all points and these centroids. The points which have the closest centroid are attributed to that cluster. Next, the center of the cluster is defined as a new centroid,

and the process repeats until the location of the centroid does not move. This is a simple linear method, and is able to assess for clearly defined clusters. The second chosen method is K-medoids (Huang, 1998); a method which words very similar to K-means. The algorithm has the same linear, iterative nature, but k-medoids does not assign clusters by centroids but by existing data-points. Because of this, variables need not be linear but can also be binary, if combined with Gower-distance (highlighted later in this appendix).

The third method chosen is DBSCAN (Ester et al., 1996), or Density-based spatial clustering of applications with noise. DBSCAN works differently than k-means and k-medoids, as it clusters based on density, not directly on distance. Because of this, for instance non-linearly shaped clusters can be identified. Also, this method works accurately when applying to data with outliers, which our dataset exhibits. Using the correct distance metric, DBSCAN can be used for both numerical and mixed datasets. Therefore, the numerical dataset is tested using K-means and DBSCAN, and the full mixed dataset is assessed using K-medoids and also DBSCAN.

A.1.2 Distance Metrics

For clustering it is also important to assess which distance metrics are used to create the clusters. Two distance metrics are used in this research, one for the numerical dataset and one for the mixed dataset. For the numerical dataset, Euclidean distance is used (Provost & Fawcett, 2013). Euclidean distance is the simple 'straight-line' distance between points in a multi-dimensional space. Euclidean distance is accurate for low-dimensional data, but as data-dimensionality increases (i.e. the number of variables of the dataset increases), it's accuracy reduces. The numerical dataset contains 8 variables, for which Euclidean distance is still deemed to be accurate enough.

For the mixed-data, Gower's distance is used (Gower, 1971). Gower's distance is computed as the average of partial dissimilarities across all data points, meaning each numerical and ordinal variable is normalized, and data-points can have a difference ranging from 0 to 1 for each variable. Categorical variables are translated into multiple binary variables (i.e., in our dataset, region would be translated into 4 boolean variables as region can take 4 values). A downside of Gower's distance is that numerical and ordinal variables have relatively less impact on the distance between data-points than boolean variables.

A.1.3 Cluster validity metrics

To assess if the methods are able to create clusters, two cluster-validity metrics are used for each test. Firstly the silhouette-score is used (Rousseeuw, 1987). The silhouette score evaluates the average distance of each datapoint to the average distance to other data-points. The score can range from 1 to -1, whereby 1 indicates datapoint which are closest to the cluster, and -1 indicates the datapoint is closest to a different cluster, indicating inaccurate clustering. The second index used is the S_Dbw validity index (Halkidi et al., 2002). This index sums up the average intra-cluster variance and the inter-cluster density to create a score. As intra-cluster variance is minimal for clear clusters, and also for inter-cluster density, a low S_Dbw index indicates better clusters.

A.1.4 Approach

To assess if clusters possibility exist in a certain part of the data, two general approaches are taken towards clustering. Firstly, a fully computational approach is done, testing the full dataset for clusters with all combinations of clustering settings possible, simply to see which combinations yield the most accurate result. Next, a more theoretical approach is also taken, whereby a selection is made of the variables which are deemed to have the largest possible impact for clusters. The primary theoretical background is to see in which aspects 'good' and 'bad' privacy statements differ. The variables were subsequently ranked, in regards to which of

the selected variables probably had the largest effect. This was done in cooperation with two Privacy advisers from the Deloitte Privacy Services team.

The following list of variables was chosen:

- Informativeness
- Vagueness
- Flesch Score
- Legitimate Interest
- Social Media
- Email Provided
- Change Notification
- After GDPR
- Data Combining
- Location Collection
- Location Sharing
- Category
- Region

A.1.5 Python

As in the rest of this research, Python is used to apply the clustering algorithms. Apart from the standard packages used in this research to manage the data (**Numpy** and **Pandas**) and visualization (**matplotlib** and **Seaborn**) a number of additional specific clustering packages are used. For K-means and DBSCAN, the **sklearn** package is used, which contains standard pre-written clustering algorithms for these methods. The metrics package from sklearn is also used to test the silhouette score. For k-medoids, the specific package **Kmodes** is used, which contains a specific package for this method. Lastly, the package **s_dbw** contains a pre-written S_dbw index.

A.2 Results

After rigorous testing of all the methods above including varying settings within the clustering alogirthms, none yielded significant, valuable results. In a few cases, the cluster validity metrics would indicate some sort of structure was found, but after visualization of the data, this was mostly due to correlation within the data, giving the impression clusters exist. Results for similar for the full dataset and numerical dataset, showing no clear structures. Applying the theoretical approach and testing for all combinations of variables within the selection also did not yield any tangible results. Because of this, results will not be discussed in further detail.

In principle, not finding clusters is not negative, as it simply indicates the finding that the data does not contain clusters. However, a problem in this specific situation is that the data is also not ideal for clustering. Although the clustering methods for mixed data exist, they also exhibit certain drawbacks (Birant & Kut, 2007), as mentioned the variables are normalized. Also, mixed data is less likely to exhibit clusters as the granularity of the data is very low (Provost & Fawcett, 2013). Therefore, when applying clustering methods

to this dataset, the chance of finding clusters is already low as the data is unlikely to exhibit clusters. This is also true for the numerical dataset, as it contains ordinal variables. Although ordinal variables have a higher chance of providing clusters than booleans, the granularity is still a lot lower than numerical variables, meaning it still reduces the possibility of clusters.

If clustering was the main focus of this research, a stronger focus would need to be put on extracting numerical variables from the privacy statements, as these further improve the chance of finding clusters. Furthermore, this would need to be done in combination with a theoretical approach, making sure the variables are extracted for which clusters can theoretically be expected. In this case, the attempt at clustering can be seen as a rigorous test, confirming that in this dataset, the presence of clusters is highly unlikely.

Appendix B

Full Regional Analysis

This appendix shows the results of all Kruskal-Wallis tests performed, with region as an independent variable and the numeric variables as dependent variables. The first five results are based on the full dataset, the last five results are based on the dataset containing statements with a maximum age of 60 months.

B.1 Kruskal Wallis Tests

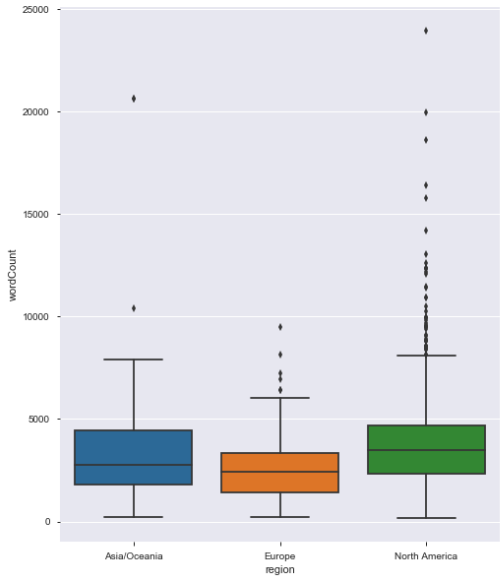
Per numeric variable, a screenshot is shown with three results. The first result is the result of the Kruskal-Wallis test, with the corresponding statistic and p-value. A p-value smaller than 0.05 indicates significant differences between regions, but does not specify which regions. Second, a table is given as a result of Dunn's test, which indicates between which regions a statistically significant value is found for that variable. The second result is a boxplot, indicating the spread of each variable per group. Lastly, a table is given as a result of Dunn's test, which indicates between which regions a statistically significant value is found for that variable.

B.2 Chi-squared contingency tests

Next, the analysis is performed for all boolean variables in the dataset. The same tests are performed on a number of different subsets of the dataset to assess the effects of modifying the data. This is done due to the initially perculiar results of North American statements having higher frequencies for nearly all of the variables. These extra tests give insight into why this effect is caused. Removing all old statements (more than 60 months of age), all uninformative statements (informative score of 5 or lower) and also all long North American statements (the 400 longest NA statements removed) all have effects on the data, and balance the data out. The results can be seen in figures at the bottom of this appendix (figures B.9, B.10, B.11 and B.12).

wordCount
KruskalResult(statistic=40.349770927542394, pvalue=1.730448066885603e-09)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	True
North America	True	True	False



wordCount
KruskalResult(statistic=23.541633609358787, pvalue=7.726791870907503e-06)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	False
Europe	True	False	True
North America	False	True	False

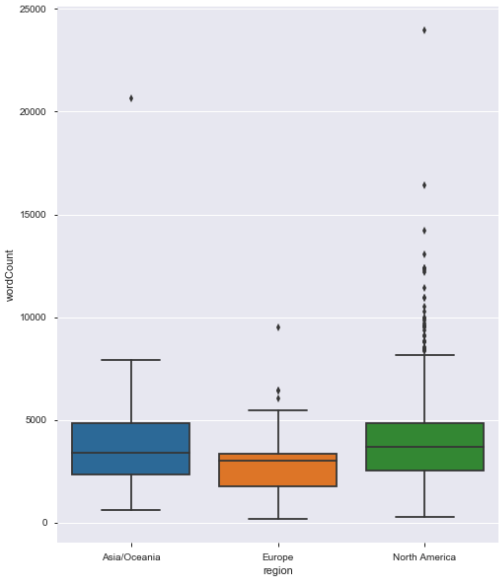
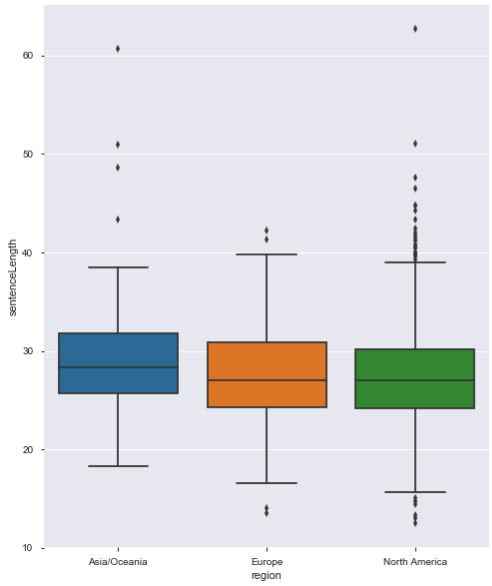


Figure B.1: Kruskall-Wallis results for Region and Word Count, Full Dataset Left, New Dataset Right

sentenceLength
KruskalResult(statistic=7.496274565001527, pvalue=0.023561593598057633)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	False
North America	True	False	False



sentenceLength
KruskalResult(statistic=8.583953053939798, pvalue=0.013677863895499544)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	False
North America	True	False	False

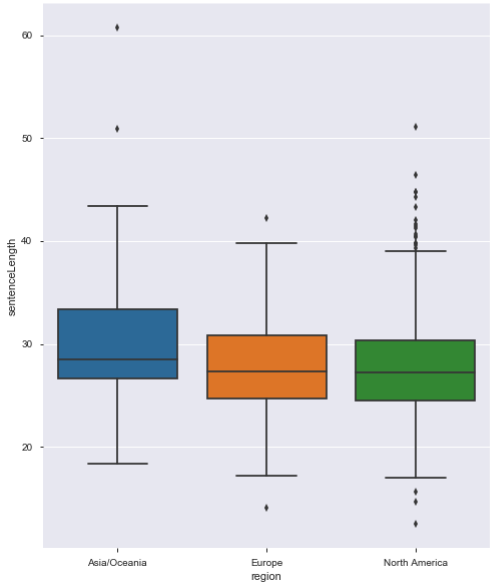
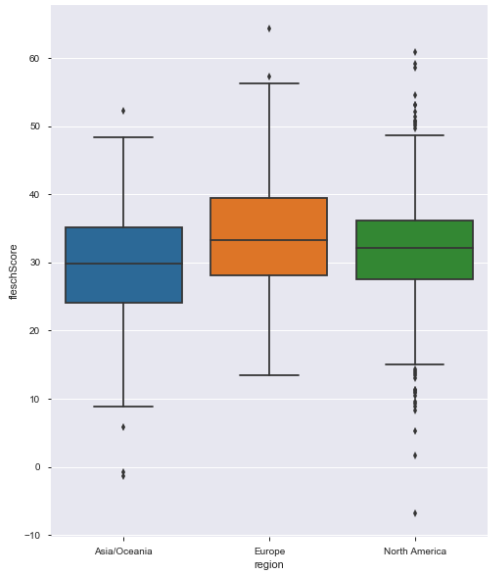


Figure B.2: Kruskall-Wallis results for Region and Sentence Length, Full Dataset Left, New Dataset Right

fleschScore
KruskalResult(statistic=13.611128036172328, pvalue=0.0011075952911386619)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	True
North America	True	True	False



fleschScore
KruskalResult(statistic=14.30014823812138, pvalue=0.0007848059100782805)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	False
North America	True	False	False

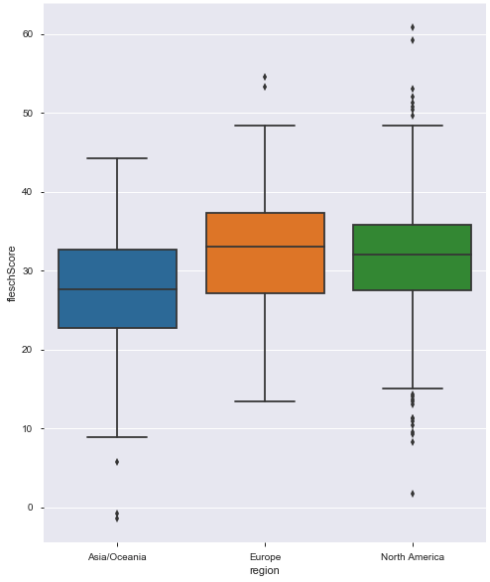
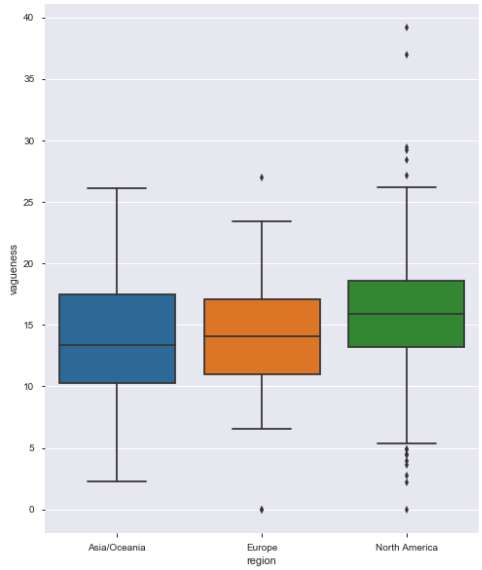


Figure B.3: Kruskall-Wallis results for Region and FRE-score, Full Dataset Left, New Dataset Right

vagueness
KruskalResult(statistic=30.886790369795044, pvalue=1.9634447695341272e-07)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	True
Europe	False	False	True
North America	True	True	False



vagueness
KruskalResult(statistic=10.450067921119892, pvalue=0.0053801772429764695)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	True
Europe	False	False	True
North America	True	True	False

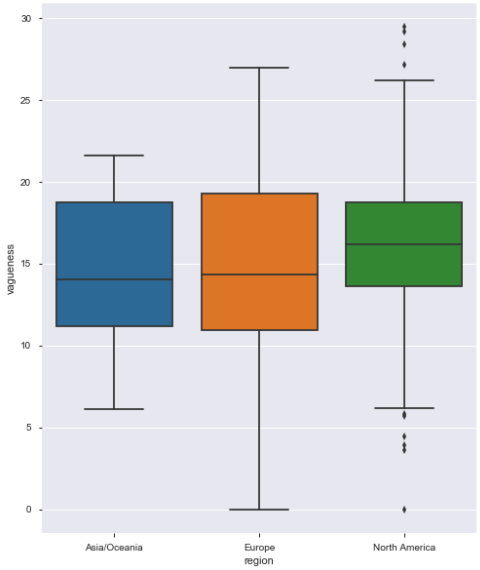
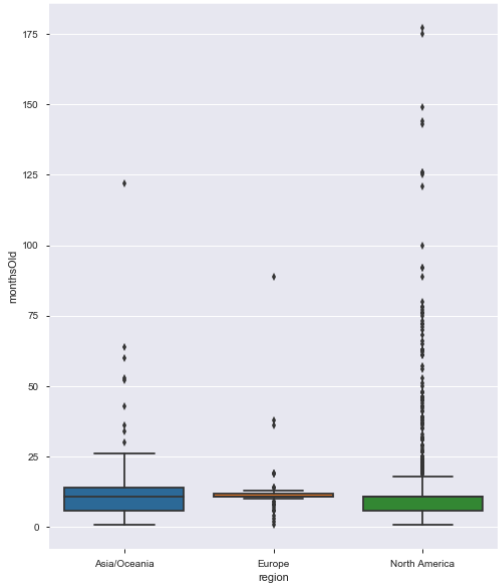


Figure B.4: Kruskall-Wallis results for Region and Vagueness, Full Dataset Left, New Dataset Right

```
monthsOld
KruskalResult(statistic=6.0081083463497995, pvalue=0.04958563157655449)
```

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	False
Europe	False	False	False
North America	False	False	False




```
monthsOld
KruskalResult(statistic=8.543553299309607, pvalue=0.013956964476700334)
```

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	False
Europe	False	False	True
North America	False	True	False

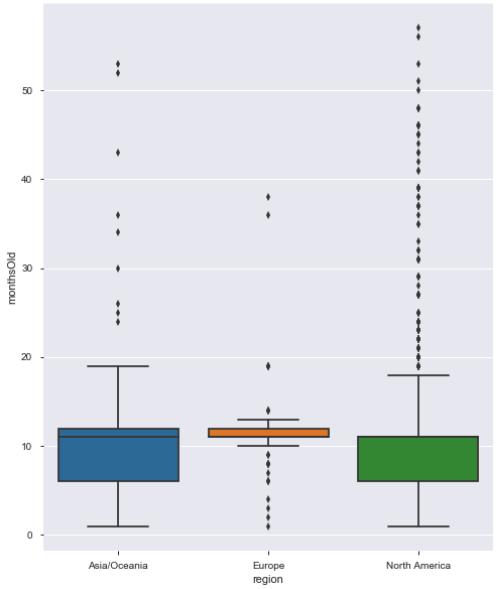
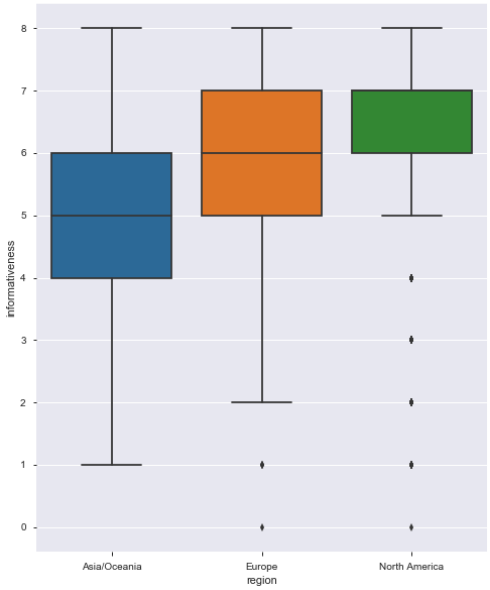


Figure B.5: Kruskall-Wallis results for Region and Age in Months, Full Dataset Left, New Dataset Right

informativeness
KruskalResult(statistic=91.1929278917251, pvalue=1.576548421388855e-20)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	True
North America	True	True	False



informativeness
KruskalResult(statistic=56.22531020872559, pvalue=6.177731325014119e-13)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	True
Europe	False	False	True
North America	True	True	False

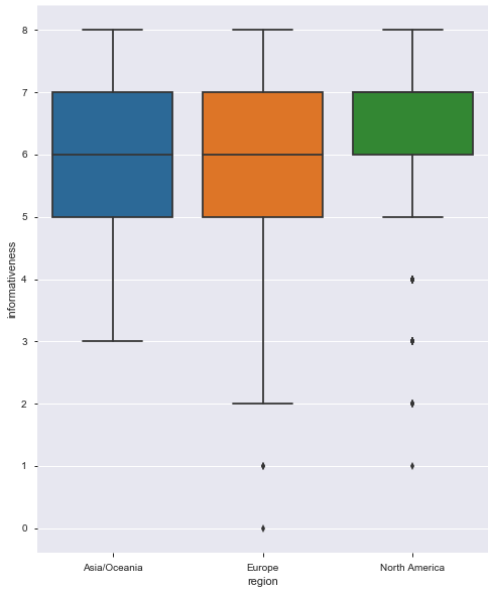
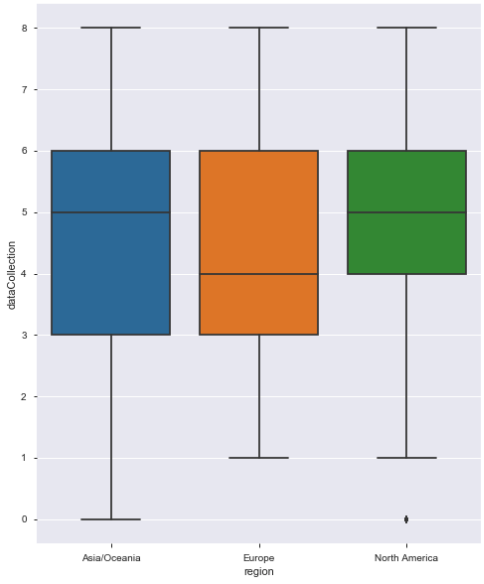


Figure B.6: Kruskall-Wallis results for Region and Informativeness, Full Dataset Left, New Dataset Right

dataCollection
KruskalResult(statistic=39.85597234904326, pvalue=2.2150603793477454e-09)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	True
Europe	False	False	True
North America	True	True	False



dataCollection
KruskalResult(statistic=18.101949867980917, pvalue=0.00011727664417879663)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	False	True
Europe	False	False	True
North America	True	True	False

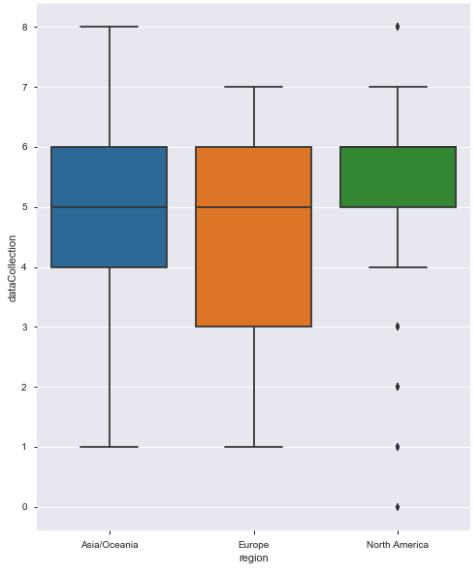
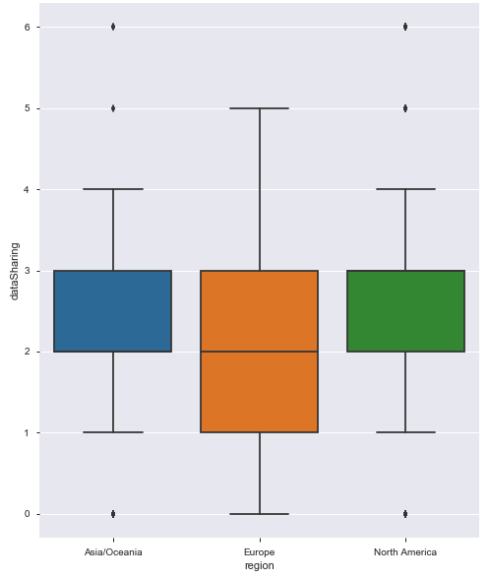


Figure B.7: Kruskall-Wallis results for Region and Data Collection, Full Dataset Left, New Dataset Right

dataSharing
KruskalResult(statistic=25.967975474975102, pvalue=2.296813713434466e-06)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	True
Europe	True	False	True
North America	True	True	False



dataSharing
KruskalResult(statistic=15.285543897137467, pvalue=0.0004794974658837432)

	Asia/Oceania	Europe	North America
Asia/Oceania	False	True	False
Europe	True	False	True
North America	False	True	False

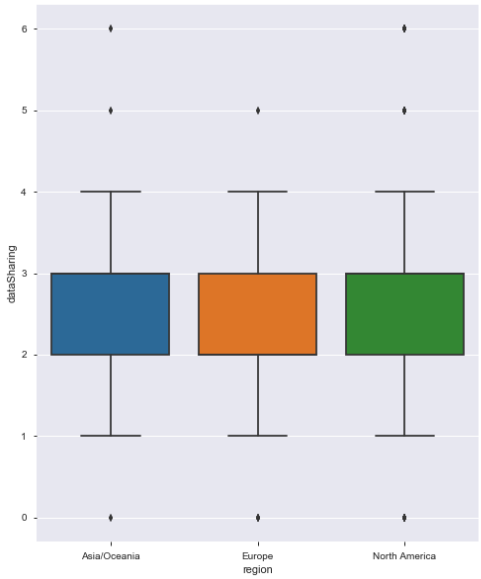


Figure B.8: Kruskall-Wallis results for Region and Data Sharing, Full Dataset Left, New Dataset Right

	North America	Europe	Asia/Oceania	Asia/Oceania	Europe	North America
legitimateInterest	46.5%	57.0%	31.1%	lower	-	-
accessRights	87.1%	82.2%	76.7%	lower	-	higher
dpo	25.4%	41.1%	17.5%	-	higher	-
afterGDPR	68.1%	55.1%	44.7%	lower	-	higher
thirdParty	98.1%	92.5%	95.1%	-	lower	higher
choices	85.2%	64.5%	52.4%	lower	lower	higher
security	94.6%	87.9%	88.3%	-	-	higher
specificAudiences	69.4%	45.8%	38.8%	lower	lower	higher
privacyShield	30.5%	13.1%	2.9%	lower	lower	higher
dataRetention	76.7%	79.4%	60.2%	lower	-	-
policyChange	92.6%	83.2%	88.3%	-	lower	higher
google	62.4%	40.2%	33.0%	lower	lower	higher
facebook	50.2%	32.7%	34.0%	lower	lower	higher
amazon	9.9%	1.9%	2.9%	-	lower	higher
adobe	15.1%	10.3%	4.9%	lower	-	higher
doubleClick	12.2%	3.7%	11.7%	-	lower	-
paypal	9.8%	5.6%	2.9%	-	-	higher
apple	15.1%	4.7%	9.7%	-	lower	higher
socialMedia	63.9%	50.5%	41.7%	lower	-	higher
nai	31.9%	27.1%	19.4%	lower	-	higher
daa	24.3%	5.6%	4.9%	lower	lower	higher
android	15.7%	4.7%	12.6%	-	lower	higher
emailProvided	81.3%	82.2%	60.2%	lower	-	higher
computerInformation	86.7%	73.8%	72.8%	lower	lower	higher
contactInformation	93.1%	86.0%	81.6%	lower	-	higher
webTracking	52.5%	28.0%	36.9%	lower	lower	higher
dataCombining	15.2%	7.5%	6.8%	-	-	higher
demographicInformation	74.1%	56.1%	71.8%	-	lower	higher
contactInformationShare	43.1%	30.8%	25.2%	lower	-	higher
personalInformationShare	83.6%	63.6%	80.6%	-	lower	higher
locationShare	14.7%	5.6%	13.6%	-	lower	-
generalShare	95.3%	86.0%	88.3%	-	lower	higher
changeNotification	33.4%	16.8%	18.4%	lower	lower	higher
addressProvided	75.2%	65.4%	47.6%	lower	-	higher

Figure B.9: Chi-squared contingency test with Bonferroni corrections, Full Dataset

	North America	Europe	Asia/Oceania	North America	Europe	Asia/Oceania
legitimateInterest	46.5%	57.0%	31.1%	lower	higher	-
dpo	25.4%	41.1%	17.5%	-	higher	-
choices	85.2%	64.5%	52.4%	higher	-	-
privacyShield	30.5%	13.1%	2.9%	higher	lower	lower
dataRetention	76.7%	79.4%	60.2%	lower	-	-
google	62.4%	40.2%	33.0%	higher	lower	-
facebook	50.2%	32.7%	34.0%	-	lower	-
amazon	9.9%	1.9%	2.9%	higher	-	-
criteo	2.0%	1.9%	4.9%	-	-	higher
apple	15.1%	4.7%	9.7%	-	lower	-
daa	24.3%	5.6%	4.9%	higher	lower	lower
webTracking	52.5%	28.0%	36.9%	higher	lower	-
personalInformationShare	83.6%	63.6%	80.6%	-	lower	-
changeNotification	33.4%	16.8%	18.4%	higher	-	-
addressProvided	75.2%	65.4%	47.6%	higher	lower	lower

Figure B.10: Chi-squared contingency test with Bonferroni corrections, Informative statements only

	North America	Europe	Asia/Oceania	North America	Europe	Asia/Oceania
legitimateInterest	46.5%	57.0%	31.1%	-	higher	-
contractualNecessity	41.4%	43.9%	38.8%	lower	-	-
dpo	25.4%	41.1%	17.5%	lower	higher	-
dataController	29.1%	39.3%	23.3%	lower	higher	-
afterGDPR	68.1%	55.1%	44.7%	higher	-	lower
choices	85.2%	64.5%	52.4%	higher	lower	lower
specificAudiences	69.4%	45.8%	38.8%	higher	lower	lower
privacyShield	30.5%	13.1%	2.9%	higher	-	lower
google	62.4%	40.2%	33.0%	higher	-	lower
amazon	9.9%	1.9%	2.9%	higher	-	-
yahoo	3.9%	1.9%	6.8%	-	-	higher
daa	24.3%	5.6%	4.9%	higher	-	-
emailProvided	81.3%	82.2%	60.2%	-	-	lower
computerInformation	86.7%	73.8%	72.8%	higher	-	-
webTracking	52.5%	28.0%	36.9%	higher	lower	-
personalInformationShare	83.6%	63.6%	80.6%	-	lower	-
generalShare	95.3%	86.0%	88.3%	higher	-	-
changeNotification	33.4%	16.8%	18.4%	higher	-	-
addressProvided	75.2%	65.4%	47.6%	higher	-	lower

Figure B.11: Chi-squared contingency test with Bonferroni corrections, 400 longest NA statements removed

	North America	Europe	Asia/Oceania	Asia/Oceania	Europe	North America
dpo	25.7%	46.2%	20.7%	-	higher	-
thirdParty	99.2%	93.8%	98.3%	-	lower	higher
choices	87.6%	72.3%	56.9%	lower	lower	higher
security	97.3%	87.7%	96.6%	-	lower	higher
specificAudiences	73.8%	49.2%	46.6%	lower	lower	higher
privacyShield	32.8%	13.8%	3.4%	lower	lower	higher
google	63.3%	38.5%	32.8%	lower	lower	higher
facebook	53.3%	33.8%	34.5%	lower	lower	higher
amazon	10.2%	1.5%	1.7%	-	-	higher
socialMedia	67.6%	53.8%	44.8%	lower	-	higher
daa	26.7%	4.6%	6.9%	lower	lower	higher
android	17.4%	3.1%	15.5%	-	lower	-
emailProvided	84.1%	87.7%	62.1%	lower	-	-
computerInformation	89.2%	80.0%	81.0%	-	-	higher
webTracking	56.3%	29.2%	44.8%	-	lower	higher
dataCombining	16.5%	7.7%	6.9%	-	-	higher
demographicInformation	76.7%	61.5%	75.9%	-	lower	-
contactInformationShare	46.1%	33.8%	34.5%	-	-	higher
personalInformationShare	86.7%	70.8%	89.7%	-	lower	-
generalShare	97.3%	90.8%	96.6%	-	lower	-
changeNotification	35.8%	15.4%	19.0%	-	lower	higher
addressProvided	77.1%	63.1%	53.4%	lower	-	higher

Figure B.12: Chi-squared contingency test with Bonferroni corrections, Statements less than 60 months old

Appendix C

Full Category Analysis

This appendix gives all results of the category analysis tests. First, the results of the Kruskal-Wallis H-tests are shown for each numeric variable. Next, the full tables are shown with all significant differences for the Chi-squared tests with Bonferroni corrections.

C.1 Kruskal-Wallis H-tests

For each numeric variable, an image is shown with all results. First, the result of the Kruskal-Wallis test with corresponding p-value is shown. A p-value of less than 0.05 indicates a significant difference somewhere between the groups. Next, a table is shown of all categories, and the amount of significant differences the category has with other categories. This is done to keep an overview of the results, a full table as used in the previous Appendix is more informative, but less informative. Next to the amount of significant differences, the mean value of the variable for that category is added, to indicate if the variable is significantly lower or higher than average. Last, the full boxplot is shown which also indicates what the values yield per category.

```
wordCount
KruskalResult(statistic=57.058950760328145, pvalue=3.820841111892631e-07)
```

	Significant Differences	Mean
Law	12	2091.230769
Technology & Computing	8	3022.400000
Personal Finance	7	2937.837209
News / Weather / Information	7	3013.038596
Travel	6	4591.648649
Style & Fashion	5	4415.636364
Non-Standard Content	5	2966.142857
Health & Fitness	5	3944.571429
Business	5	3651.231405
Arts & Entertainment	5	3659.739726
Shopping	4	3938.354839
Sports	3	3489.500000
Education	2	3350.072464
Society	1	3480.030303
Hobbies & Interests	1	3447.348485

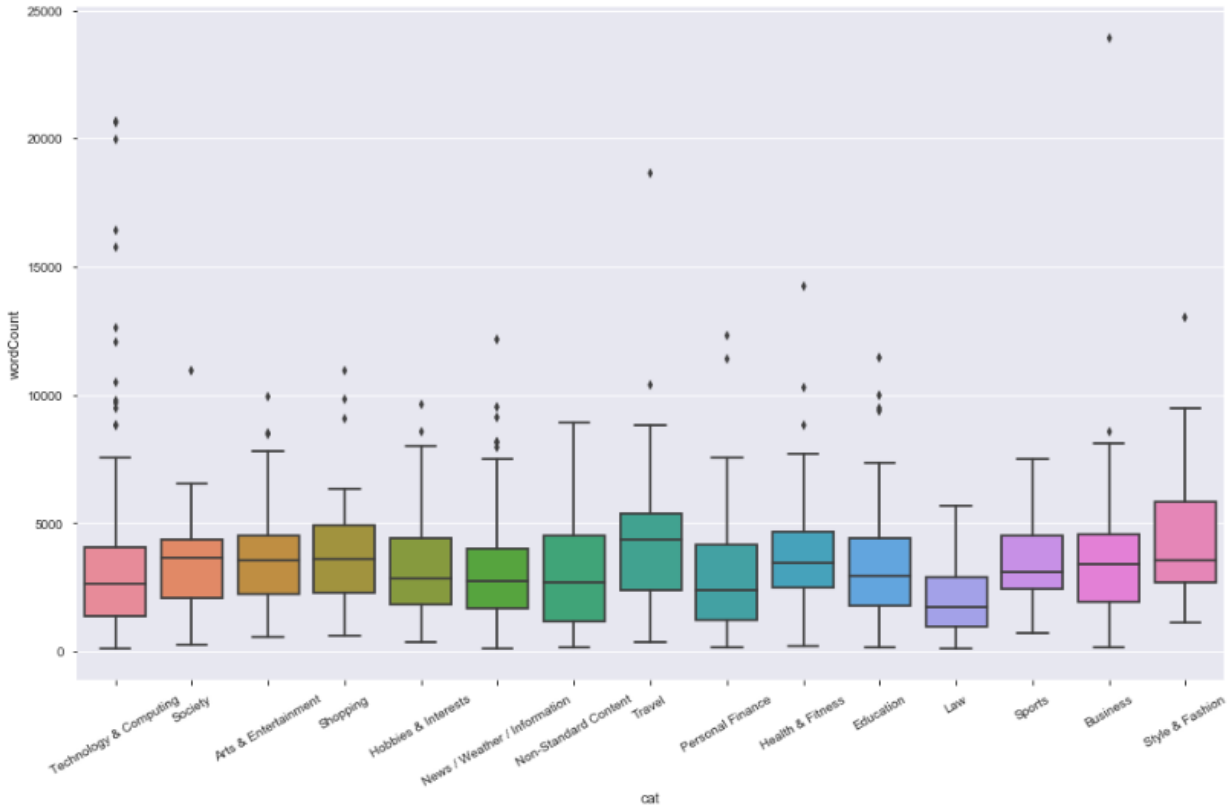


Figure C.1: Kruskall-Wallis results for Category and Word Count

```
sentenceLength
KruskalResult(statistic=33.71752489636255, pvalue=0.0022676746631192022)
```

	Significant Differences	Mean
Sports	8	29.470183
Law	6	25.345059
Technology & Computing	5	26.524761
Non-Standard Content	5	25.768262
Style & Fashion	4	29.078157
Personal Finance	4	28.056059
Health & Fitness	3	27.912291
Education	3	26.270104
Arts & Entertainment	3	27.920760
Travel	1	28.272688
Society	1	26.657803
News / Weather / Information	1	26.898935
Hobbies & Interests	1	26.745333
Business	1	26.936719
Shopping	0	27.428573

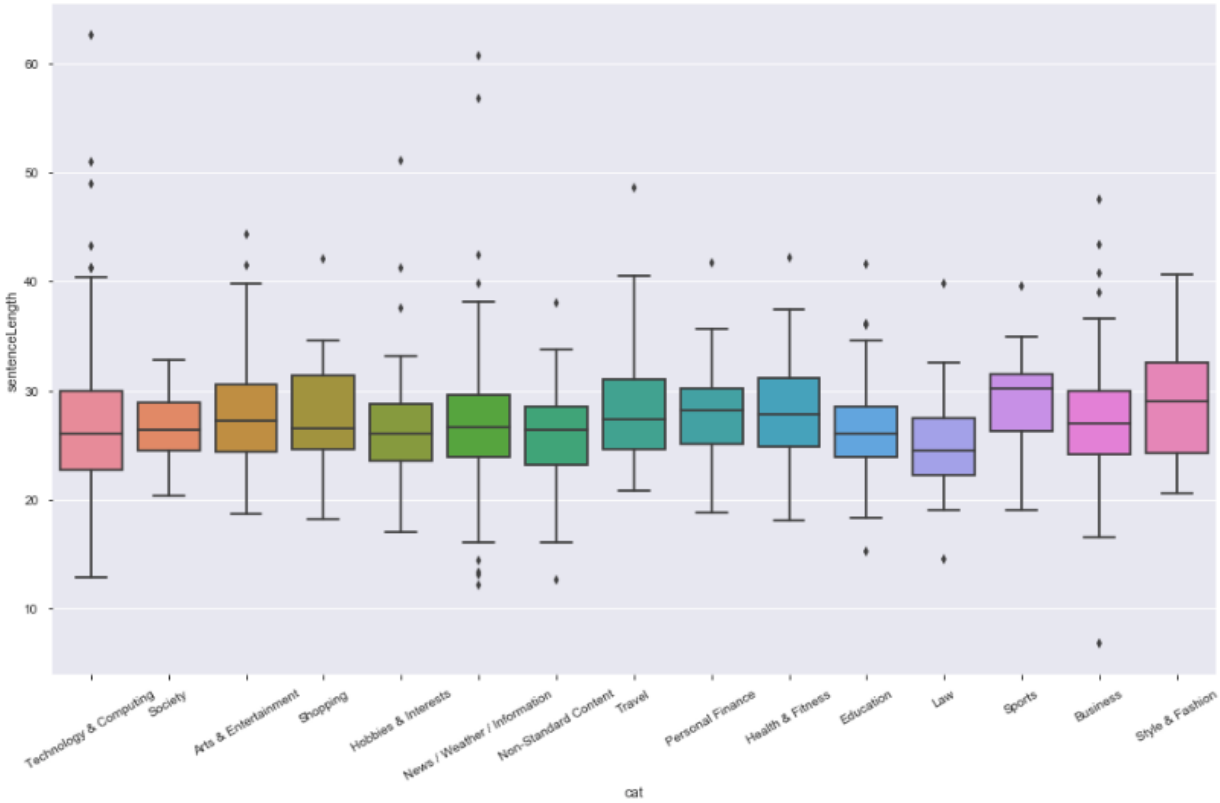


Figure C.2: Kruskall-Wallis results for Category and Sentence Length

```
fleschScore
KruskalResult(statistic=25.252540571997205, pvalue=0.03215457672802549)
```

	Significant Differences	Mean
Sports	9	28.729240
Society	7	34.862300
Business	2	31.976220
Arts & Entertainment	2	31.698550
Travel	1	30.701098
Technology & Computing	1	32.873728
Shopping	1	31.039766
Personal Finance	1	31.884268
Non-Standard Content	1	34.079977
News / Weather / Information	1	32.586460
Law	1	34.313568
Hobbies & Interests	1	32.918793
Health & Fitness	1	31.575490
Education	1	33.262426
Style & Fashion	0	32.185943

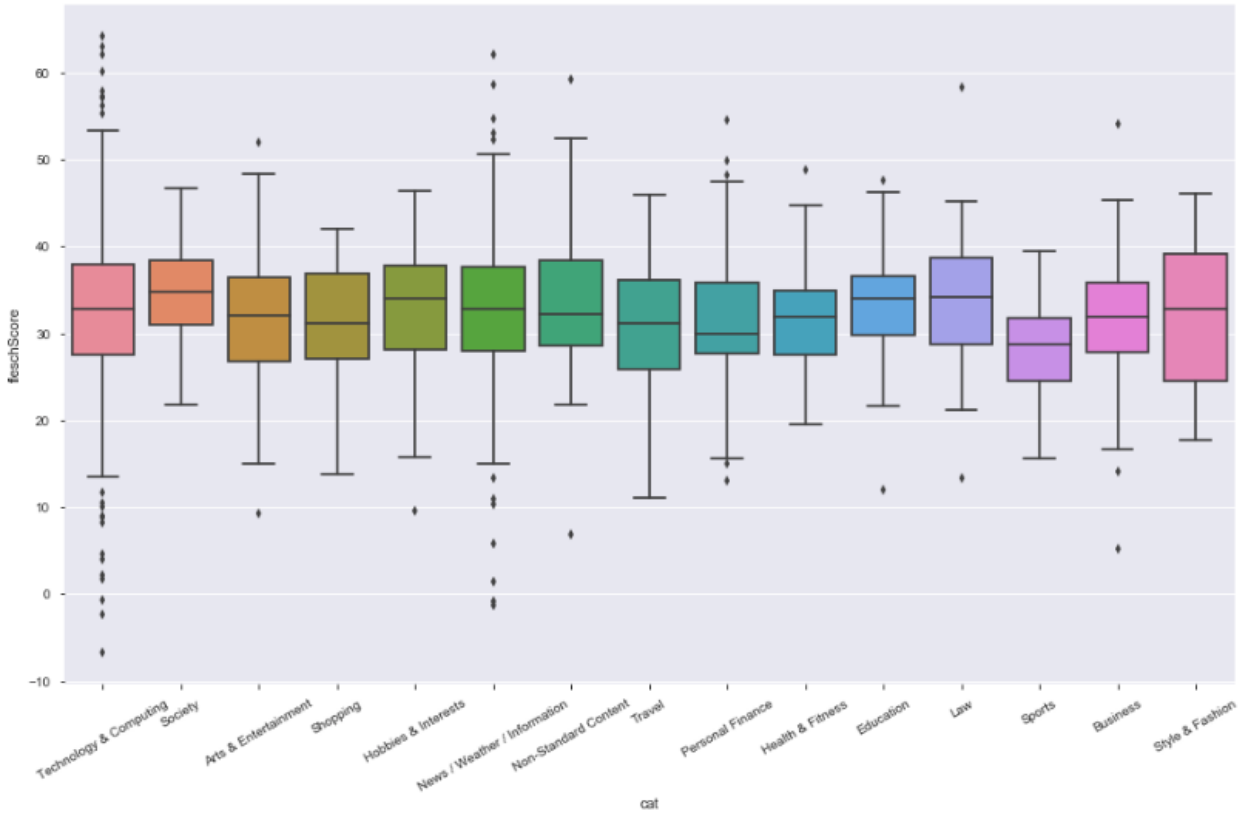


Figure C.3: Kruskall-Wallis results for Category and Flesch Score

```
vagueness
KruskalResult(statistic=42.02827569195319, pvalue=0.00012235791485871937)
```

	Significant Differences	Mean
Education	7	13.985253
Business	7	14.450768
Technology & Computing	5	14.899266
Health & Fitness	5	16.687068
Arts & Entertainment	5	16.669353
Personal Finance	3	16.362372
News / Weather / Information	3	15.525898
Hobbies & Interests	3	16.137880
Sports	2	16.463065
Society	2	14.847791
Shopping	2	14.839367
Non-Standard Content	2	15.764655
Travel	0	15.410503
Style & Fashion	0	16.112182
Law	0	15.543709

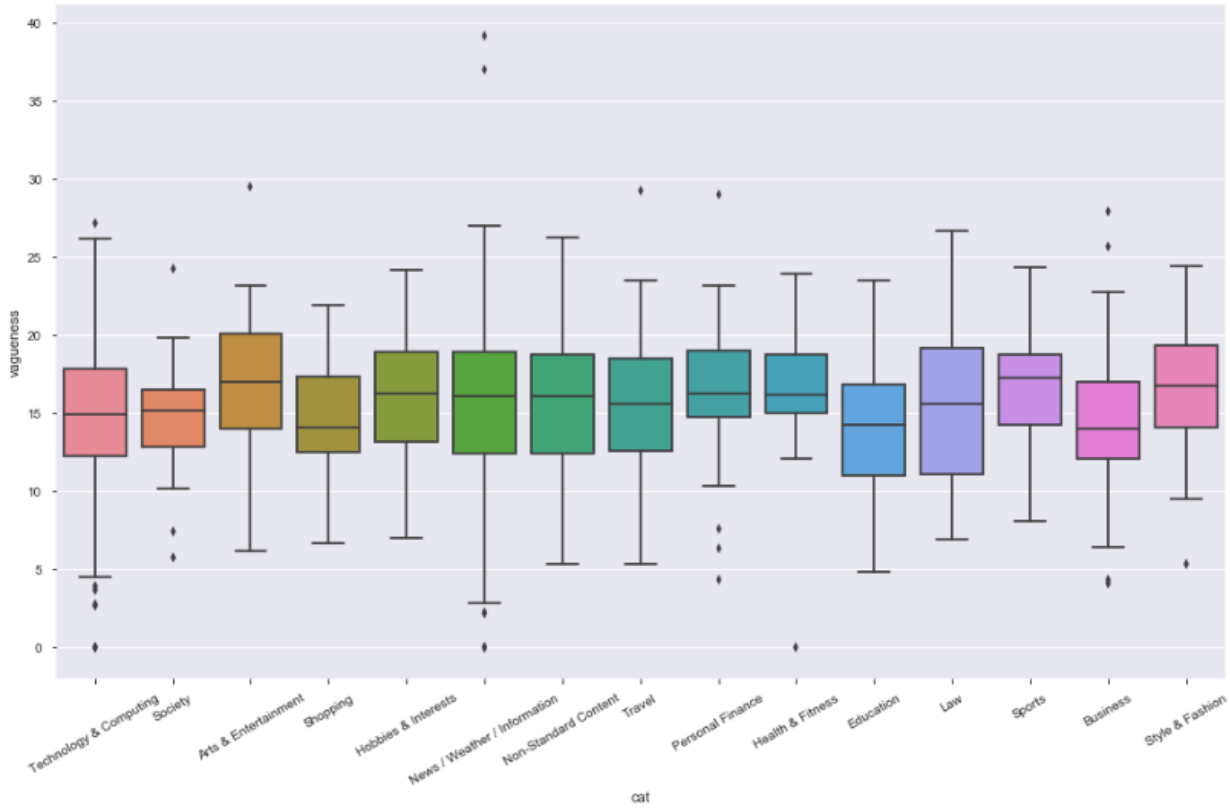


Figure C.4: Kruskall-Wallis results for Category and Vagueness

```
monthsOld
KruskalResult(statistic=22.92190288091364, pvalue=0.061552436144236784)
```

	Significant Differences	Mean
Health & Fitness	2	10.000000
Society	1	14.269231
News / Weather / Information	1	17.563452
Travel	0	17.148148
Technology & Computing	0	14.605016
Style & Fashion	0	20.941176
Sports	0	17.130435
Shopping	0	18.120000
Personal Finance	0	19.875000
Non-Standard Content	0	24.520000
Law	0	30.200000
Hobbies & Interests	0	19.277778
Education	0	13.080000
Business	0	13.440860
Arts & Entertainment	0	13.476190

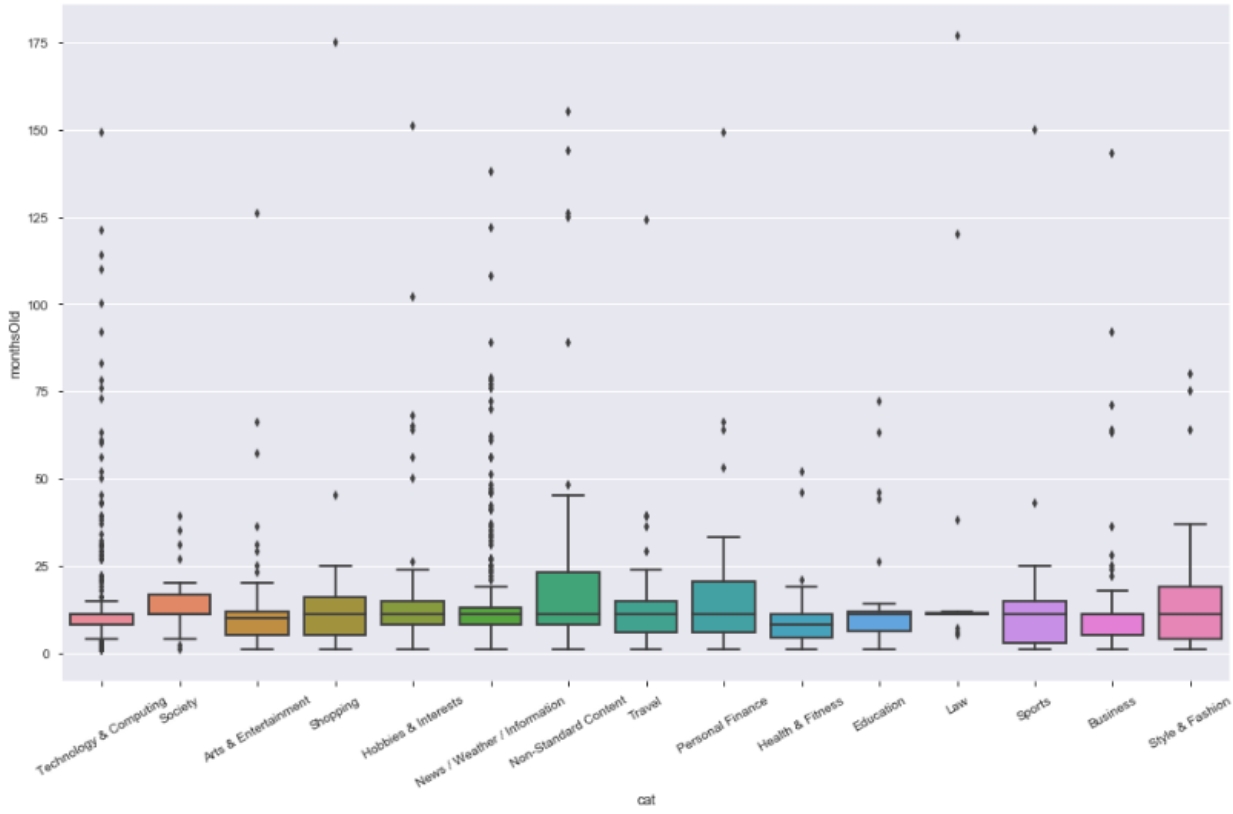


Figure C.5: Kruskall-Wallis results for Category and Months Old

```
informativeness
KruskalResult(statistic=43.416987004401605, pvalue=7.345863811071593e-05)
```

	Significant Differences	Mean
News / Weather / Information	7	5.568421
Technology & Computing	5	5.685106
Law	5	5.115385
Health & Fitness	5	6.408163
Business	5	6.231405
Personal Finance	4	5.418605
Hobbies & Interests	4	6.348485
Arts & Entertainment	4	6.301370
Style & Fashion	2	6.545455
Non-Standard Content	2	5.682540
Education	1	5.956522
Travel	0	6.135135
Sports	0	6.107143
Society	0	6.000000
Shopping	0	6.258065

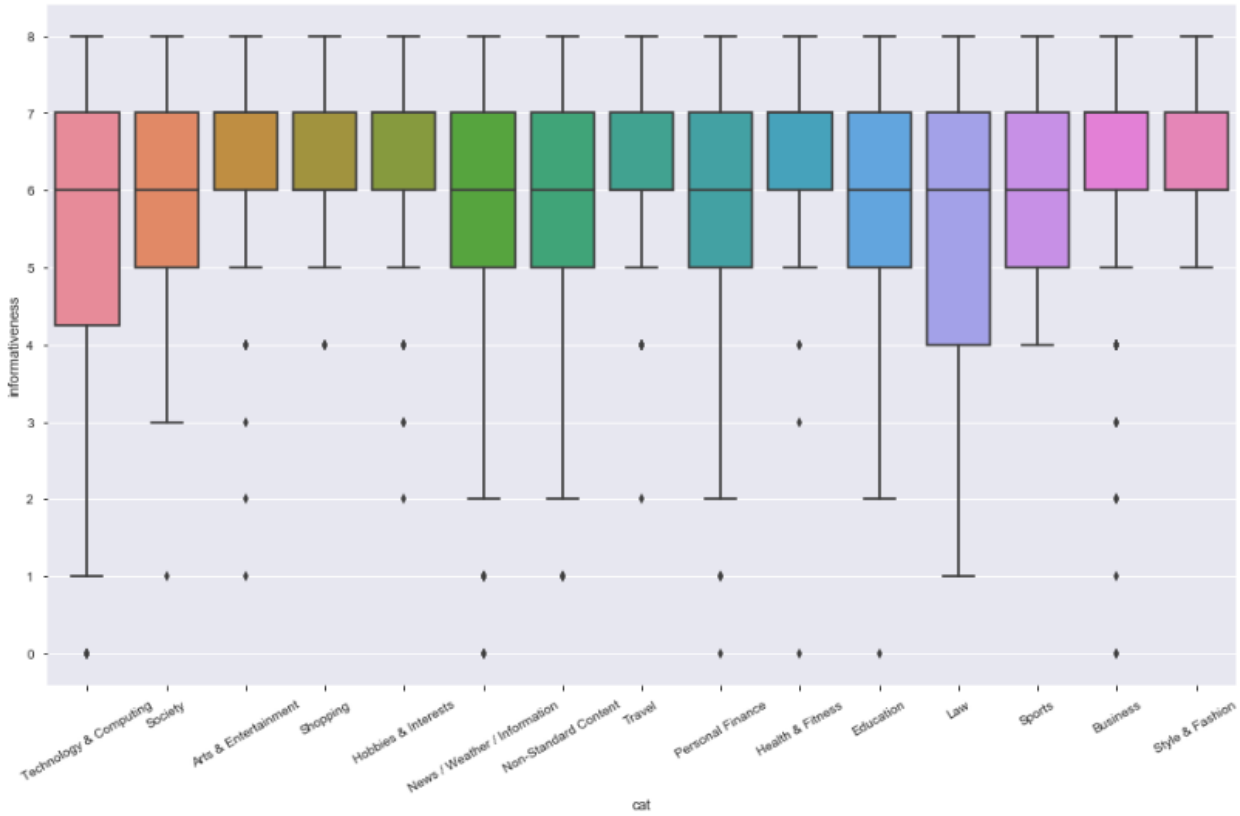


Figure C.6: Kruskall-Wallis results for Category and Informativeness

```
dataCollection
KruskalResult(statistic=43.829770901843254, pvalue=6.305476654828574e-05)
```

	Significant Differences	Mean
Health & Fitness	10	5.632653
Travel	6	5.270270
Style & Fashion	6	5.500000
Technology & Computing	5	4.442553
Personal Finance	5	4.209302
Law	5	4.038462
Arts & Entertainment	5	5.205479
Non-Standard Content	4	4.460317
News / Weather / Information	4	4.666667
Society	3	4.636364
Shopping	3	5.161290
Sports	1	4.857143
Hobbies & Interests	1	4.924242
Education	1	4.797101
Business	1	4.752066

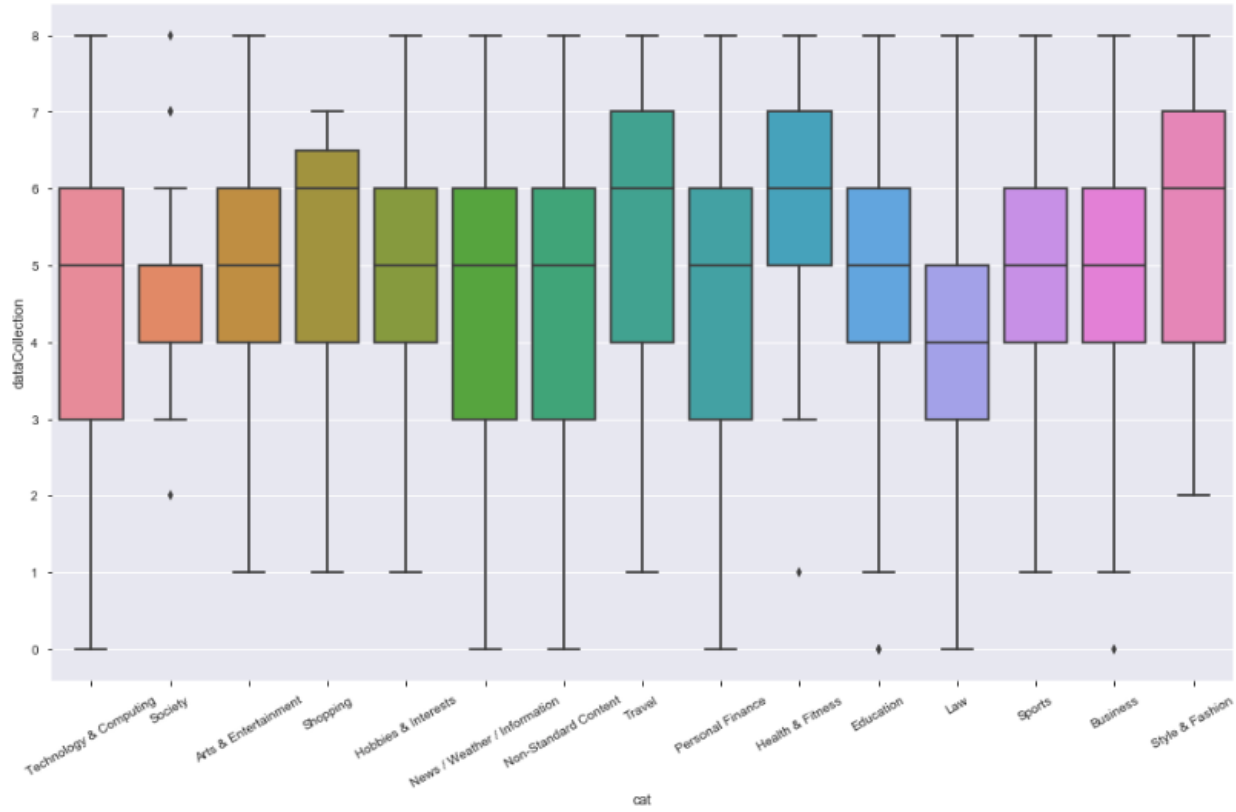


Figure C.7: Kruskall-Wallis results for Category and Data Collection

```
dataSharing
KruskalResult(statistic=78.67057680652292, pvalue=4.988279849834568e-11)
```

	Significant Differences	Mean
Law	11	1.730769
Technology & Computing	9	2.140426
Education	9	2.086957
Travel	8	3.054054
Style & Fashion	8	3.227273
Shopping	7	3.064516
News / Weather / Information	7	2.487719
Arts & Entertainment	7	2.863014
Business	6	2.363636
Personal Finance	5	2.093023
Non-Standard Content	5	2.365079
Hobbies & Interests	5	2.484848
Health & Fitness	5	2.734694
Sports	3	2.750000
Society	3	2.696970

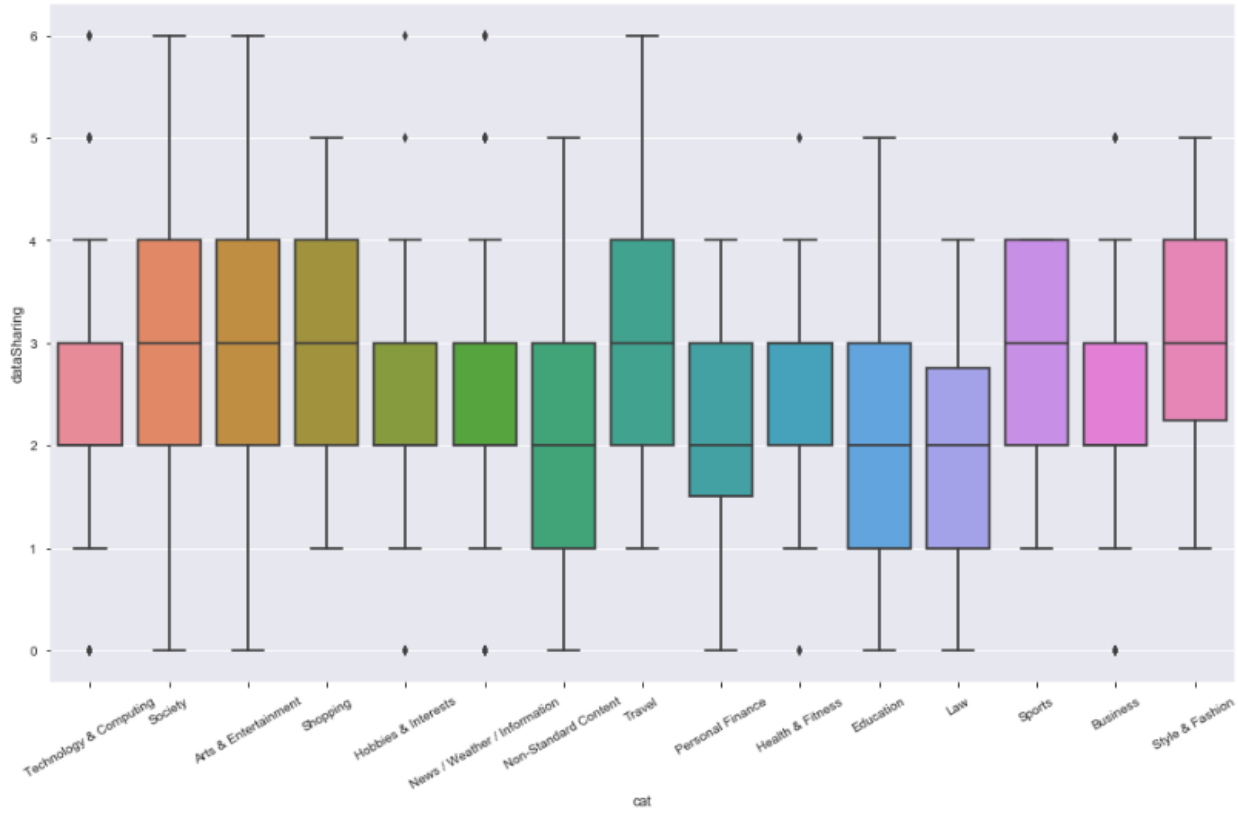


Figure C.8: Kruskall-Wallis results for Category and Data Sharing

C.2 Chi-squared contingency tests

As in the previous appendix, figures C.9 & C.10 show the results for the Chi-squared tests with Bonferroni corrections, for the full and the new dataset. Each table shows the true-percentage of the variable, together with the significant difference (where '-' indicates no significant difference). The new dataset yields less significant results, but nevertheless similar results. As the Kruskal-Wallis test did not provide significant results for age and the for these tests provide similar, but less results, it can be expected that age has no significant impact for these tests.

	Technology & Computing	Arts & Entertainment	Shopping	News / Weather / Information	Travel	Health & Fitness	Sports	Business	Average
vitalInterest	5.7% , -	4.1% , -	9.7% , -	1.8% , lower	21.6% , higher	12.2% , -	3.6% , -	9.9% , -	5.9%
legitimateInterest	40.4% , -	38.4% , -	35.5% , -	31.6% , lower	48.6% , -	46.9% , -	50.0% , -	54.5% , higher	40.5%
contractualNecessity	38.1% , -	32.9% , -	38.7% , -	26.7% , lower	62.2% , higher	44.9% , -	53.6% , -	47.1% , -	36.4%
legalObligations	28.5% , -	32.9% , -	29.0% , -	17.5% , lower	45.9% , -	32.7% , -	39.3% , -	36.4% , -	29.5%
dpo	22.1% , -	20.5% , -	25.8% , -	15.8% , lower	37.8% , -	22.4% , -	14.3% , -	28.9% , -	23.3%
dataController	29.4% , -	23.3% , -	25.8% , -	17.5% , lower	27.0% , -	22.4% , -	42.9% , -	29.8% , -	25.6%
dataProcessor	15.5% , higher	5.5% , -	3.2% , -	4.2% , lower	2.7% , -	0.0% , -	0.0% , -	15.7% , -	9.5%
thirdParty	91.9% , lower	95.9% , -	100.0% , -	95.4% , -	94.6% , -	98.0% , -	96.4% , -	96.7% , -	94.6%
specificAudiences	50.0% , lower	75.3% , -	64.5% , -	60.4% , -	59.5% , -	81.6% , higher	64.3% , -	57.9% , -	58.7%
privacyShield	26.8% , -	20.5% , -	19.4% , -	12.3% , lower	24.3% , -	16.3% , -	10.7% , -	40.5% , higher	22.9%
dataRetention	70.6% , -	68.5% , -	80.6% , -	51.9% , lower	83.8% , -	69.4% , -	78.6% , -	83.5% , higher	68.6%
microsoft	6.8% , -	12.3% , -	0.0% , -	9.5% , -	0.0% , -	8.2% , -	25.0% , higher	6.6% , -	8.0%
adobe	5.7% , lower	19.2% , -	29.0% , -	17.2% , -	13.5% , -	14.3% , -	39.3% , higher	13.2% , -	12.2%
doubleClick	7.9% , -	13.7% , -	6.5% , -	16.8% , higher	8.1% , -	14.3% , -	3.6% , -	9.1% , -	10.7%
nielsen	0.9% , lower	12.3% , higher	3.2% , -	9.8% , higher	2.7% , -	8.2% , -	7.1% , -	1.7% , -	4.7%
paypal	11.5% , higher	9.6% , -	9.7% , -	4.2% , -	5.4% , -	2.0% , -	3.6% , -	5.8% , -	8.0%
apple	8.1% , -	24.7% , higher	12.9% , -	14.4% , -	10.8% , -	14.3% , -	28.6% , -	7.4% , -	11.8%
stripe	8.3% , higher	6.8% , -	0.0% , -	1.4% , lower	2.7% , -	4.1% , -	3.6% , -	5.8% , -	5.0%
socialMedia	46.8% , lower	63.0% , -	61.3% , -	56.8% , -	70.3% , -	71.4% , -	64.3% , -	56.2% , -	53.7%
nai	19.1% , lower	42.5% , higher	29.0% , -	31.6% , -	37.8% , -	36.7% , -	21.4% , -	25.6% , -	27.1%
daa	12.8% , lower	32.9% , higher	16.1% , -	22.8% , -	13.5% , -	32.7% , -	17.9% , -	16.5% , -	18.6%
personalInformation	87.7% , lower	97.3% , -	96.8% , -	92.3% , -	100.0% , -	100.0% , -	96.4% , -	95.0% , -	91.8%
location	43.8% , lower	68.5% , -	58.1% , -	53.3% , -	62.2% , -	61.2% , -	78.6% , -	47.1% , -	52.1%
demographicInformation	59.4% , lower	68.5% , -	71.0% , -	73.3% , -	73.0% , -	79.6% , -	71.4% , -	64.5% , -	66.6%
contactInformationShare	29.1% , lower	47.9% , -	48.4% , -	37.2% , -	56.8% , -	53.1% , -	46.4% , -	28.9% , -	36.2%
webTrackingShare	3.2% , -	5.5% , -	16.1% , higher	6.0% , -	2.7% , -	10.2% , -	0.0% , -	2.5% , -	4.2%
financialInformationShare	5.7% , -	8.2% , -	29.0% , higher	10.9% , -	16.2% , -	8.2% , -	3.6% , -	7.4% , -	8.4%
generalShare	85.3% , lower	97.3% , -	100.0% , -	87.4% , -	100.0% , -	93.9% , -	100.0% , -	92.6% , -	89.3%
addressProvided	55.5% , lower	57.5% , -	61.3% , -	62.1% , -	73.0% , -	71.4% , -	64.3% , -	73.6% , -	61.2%

Figure C.9: Chi-squared contingency test with Bonferroni corrections, Full Dataset

	Technology & Computing	Shopping	Arts & Entertainment	News / Weather / Information	Sports	Business	Average
legalObligations	28.5% , -	29.0% , -	32.9% , -	17.5% , lower	39.3% , -	36.4% , -	29.5%
dpo	22.1% , -	25.8% , -	20.5% , -	15.8% , lower	14.3% , -	28.9% , -	23.3%
dataProcessor	15.5% , higher	3.2% , -	5.5% , -	4.2% , -	0.0% , -	15.7% , -	9.5%
specificAudiences	50.0% , lower	64.5% , -	75.3% , -	60.4% , -	64.3% , -	57.9% , -	58.7%
privacyShield	26.8% , -	19.4% , -	20.5% , -	12.3% , lower	10.7% , -	40.5% , higher	22.9%
dataRetention	70.6% , -	80.6% , -	68.5% , -	51.9% , lower	78.6% , -	83.5% , -	68.6%
adobe	5.7% , lower	29.0% , -	19.2% , -	17.2% , higher	39.3% , higher	13.2% , -	12.2%
doubleClick	7.9% , -	6.5% , -	13.7% , -	16.8% , higher	3.6% , -	9.1% , -	10.7%
nielsen	0.9% , lower	3.2% , -	12.3% , -	9.8% , higher	7.1% , -	1.7% , -	4.7%
apple	8.1% , -	12.9% , -	24.7% , higher	14.4% , -	28.6% , -	7.4% , -	11.8%
nai	19.1% , lower	29.0% , -	42.5% , -	31.6% , -	21.4% , -	25.6% , -	27.1%
daa	12.8% , lower	16.1% , -	32.9% , -	22.8% , higher	17.9% , -	16.5% , -	18.6%
location	43.8% , lower	58.1% , -	68.5% , -	53.3% , -	78.6% , -	47.1% , -	52.1%
demographicInformation	59.4% , -	71.0% , -	68.5% , -	73.3% , higher	71.4% , -	64.5% , -	66.6%
contactInformationShare	29.1% , lower	48.4% , -	47.9% , -	37.2% , -	46.4% , -	28.9% , -	36.2%
financialInformationShare	5.7% , -	29.0% , higher	8.2% , -	10.9% , -	3.6% , -	7.4% , -	8.4%

Figure C.10: Chi-squared contingency test with Bonferroni corrections, Full Dataset

Appendix D

Company Analysis

This Appendix shows the results for all variables for the Big-5 tech companies. In section 5.2.4 the most relevant of these variables were highlighted, this appendix shows all variables.

	Google	Facebook	Amazon	Microsoft	Apple
wordCount	7153	4300	2639	2791	4009
sentenceLength	25.91	24.85	23.777	25.14	26.90
fleschScore	39.82	42.03	34.44	33.99	29.37
fleschLevel	difficult	difficult	difficult	difficult	very_confusing
vagueness	20.55	15.11	10.98	8.24	20.45
conditionality	1.81	0.93	0.75	0.35	2.49
generalization	0.69	0.69	1.13	0.0	0.0
modality	19.85	14.41	9.85	8.24	20.45
numericQuantifier	3.21	0.69	9.09	2.50	3.74
vitalInterest	False	False	False	False	False
legitimateInterest	False	False	False	False	True
publicInterest	False	False	False	False	True
contractualNecessity	False	True	False	True	False
legalObligations	False	False	False	True	False
unambiguousConsent	False	False	False	False	False
accessRights	True	True	False	True	True
dpo	False	False	False	True	True
dataController	False	False	False	False	False
dataProcessor	False	False	False	False	False
afterGDPR	True	True	False	False	False
privacyShield	True	False	True	True	False
google	True	False	False	False	False
facebook	False	True	False	False	False
amazon	False	False	True	False	False
microsoft	False	False	False	True	False
criteo	False	False	False	False	False
adobe	False	False	False	False	False
doubleClick	False	False	False	False	False
nielsen	False	False	False	False	False
paypal	False	False	False	False	False
apple	False	False	False	False	True
stripe	False	False	False	False	False
yahoo	False	False	False	False	False
socialMedia	False	False	False	False	True
nai	False	False	True	False	False
daa	False	False	False	False	False
android	True	False	False	False	False
emailProvided	False	False	False	False	False
TLD	.com	.com	.com	.com	.com
cat	Technology & Computing	Society	Shopping	Technology & Computing	Technology & Computing
changeNotification	True	False	True	False	False
headquarters	United States	United States	United States	United States	United States
addressProvided	False	True	False	False	True
monthsOld	3.0	12.0	20.0	2	4
informativeness	7	5	6	7	7
dataCollection	6	4	5	5	7
dataSharing	4	3	1	2	5
region	North America	North America	North America	North America	North America

Appendix E

Data Report

This Appendix provides an overview of the dataset achieved from using NLP. The dataset consists of 1507 privacy statements, and 72 variables. First, image E.1 shows the positive and negative rates for all boolean variables. Next, the Python package **Pandas Profiling** is used to provide a clear overview of all numerical variables.

	True	False	Positive Rate	Negative Rate
vitalInterest	89	1418	5.91%	94.09%
legitimateInterest	611	896	40.54%	59.46%
publicInterest	399	1108	26.48%	73.52%
contractualNecessity	549	958	36.43%	63.57%
legalObligations	444	1063	29.46%	70.54%
unambiguousConsent	259	1248	17.19%	82.81%
accessRights	1162	345	77.11%	22.89%
dpo	351	1156	23.29%	76.71%
dataController	386	1121	25.61%	74.39%
dataProcessor	143	1364	9.49%	90.51%
afterGDPR	883	624	58.59%	41.41%
thirdParty	1425	82	94.56%	5.44%
choices	1141	366	75.71%	24.29%
security	1343	164	89.12%	10.88%
specificAudiences	884	623	58.66%	41.34%
privacyShield	345	1162	22.89%	77.11%
dataRetention	1034	473	68.61%	31.39%
policyChange	1308	199	86.79%	13.21%
google	819	688	54.35%	45.65%
facebook	640	867	42.47%	57.53%
amazon	116	1391	7.7%	92.3%
microsoft	121	1386	8.03%	91.97%
criteo	29	1478	1.92%	98.08%
adobe	184	1323	12.21%	87.79%
doubleClick	162	1345	10.75%	89.25%
nielsen	71	1436	4.71%	95.29%
paypal	121	1386	8.03%	91.97%
apple	178	1329	11.81%	88.19%

	True	False	Positive Rate	Negative Rate
apple	178	1329	11.81%	88.19%
stripe	75	1432	4.98%	95.02%
yahoo	53	1454	3.52%	96.48%
socialMedia	809	698	53.68%	46.32%
nai	408	1099	27.07%	72.93%
daa	281	1226	18.65%	81.35%
android	191	1316	12.67%	87.33%
emailProvided	1118	389	74.19%	25.81%
computerInformation	1198	309	79.5%	20.5%
contactInformation	1300	207	86.26%	13.74%
webTracking	682	825	45.26%	54.74%
personalInformation	1383	124	91.77%	8.23%
dataCombining	180	1327	11.94%	88.06%
financialInformation	559	948	37.09%	62.91%
location	785	722	52.09%	47.91%
demographicInformation	1003	504	66.56%	33.44%
computerInformationShare	126	1381	8.36%	91.64%
contactInformationShare	545	962	36.16%	63.84%
webTrackingShare	63	1444	4.18%	95.82%
personalInformationShare	1137	370	75.45%	24.55%
financialInformationShare	127	1380	8.43%	91.57%
locationShare	179	1328	11.88%	88.12%
demographicInformationShare	89	1418	5.91%	94.09%
generalShare	1346	161	89.32%	10.68%
changeNotification	408	1099	27.07%	72.93%
addressProvided	922	585	61.18%	38.82%
noRegion	239	1268	15.86%	84.14%

Figure E.1: Values of Boolean Variables

Overview

Dataset info

Number of variables	16
Number of observations	1507
Missing cells	412 (< 0.1%)
Duplicate rows	0 (0.0%)
Total size in memory	188.5 KiB
Average record size in memory	128.1 B

Variables types

Numeric	11
Categorical	5
Boolean	0
Date	0
URL	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

- `conditionality` has 185 (12.3%) zeros — Zeros
- `dataCollection` has 38 (< 0.1%) zeros — Zeros
- `dataSharing` has 155 (10.3%) zeros — Zeros
- `monthsOld` has 411 (27.3%) missing values — Missing
- `numericQuantifier` has 91 (6.0%) zeros — Zeros

cat
Categorical

Distinct count	26
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Frequency | Composition

Value	Count	Frequency (%)
Technology & Computing	471	31.3%
News / Weather / Information	285	18.9%
Business	121	8.0%
Arts & Entertainment	73	< 0.1%
Education	69	< 0.1%
Hobbies & Interests	66	< 0.1%
Non-Standard Content	63	< 0.1%
Health & Fitness	49	< 0.1%
Personal Finance	43	< 0.1%
Travel	37	< 0.1%
Other values (16)	230	15.3%

conditionality
Numeric

Distinct count	1260	**Mean**	1.5469
Unique (%)	83.6%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	8.0863
Missing (n)	0	**Zeros (%)**	12.3%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	0.72741
Median	1.355
Q3	2.1765
95-th percentile	3.6729
Maximum	8.0863
Range	8.0863
Interquartile range	1.4491

Descriptive statistics

Standard deviation	1.1928
Coef of variation	0.7711
Kurtosis	2.9206
Mean	1.5469
MAD	0.90655
Skewness	1.2683
Sum	2331.2
Variance	1.4229
Memory size	11.9 KiB

dataCollection
Numeric

Distinct count	9	**Mean**	4.7047
Unique (%)	< 0.1%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	8
Missing (n)	0	**Zeros (%)**	< 0.1%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	1
Q1	4
Median	5
Q3	6
95-th percentile	7
Maximum	8
Range	8
Interquartile range	2

Descriptive statistics

Standard deviation	1.8734
Coef of variation	0.3982
Kurtosis	-0.23702
Mean	4.7047
MAD	1.52
Skewness	-0.58508
Sum	7090
Variance	3.5097
Memory size	11.9 KiB

dataSharing
Numeric

Distinct count	7	**Mean**	2.3968
Unique (%)	< 0.1%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	6
Missing (n)	0	**Zeros (%)**	10.3%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	2
Median	2
Q3	3
95-th percentile	5
Maximum	6
Range	6
Interquartile range	1

Descriptive statistics

Standard deviation	1.303
Coef of variation	0.54362
Kurtosis	-0.029895
Mean	2.3968
MAD	1.0359
Skewness	0.091445
Sum	3612
Variance	1.6977
Memory size	11.9 KiB

fleschLevel
Categorical

Distinct count	4
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Frequency | Composition

Value	Count	Frequency (%)
difficult	933	61.9%
very_confusing	538	35.7%
fairly_difficult	30	< 0.1%
standard	6	< 0.1%

fleschScore
Numeric

Distinct count	1502	**Mean**	32.515
Unique (%)	> 99.9%	**Minimum**	-6.7424
Missing (%)	0.0%	**Maximum**	64.265
Missing (n)	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	-6.7424
5-th percentile	20.057
Q1	27.624
Median	32.608
Q3	37.45
95-th percentile	45.849
Maximum	64.265
Range	71.007
Interquartile range	9.8253

Descriptive statistics

Standard deviation	8.3667
Coef of variation	0.25732
Kurtosis	1.9749
Mean	32.515
MAD	6.2198
Skewness	-0.18229
Sum	49000
Variance	70.002
Memory size	11.9 KiB

headquarters

Categorical

headquarters

Distinct count	42
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Frequency | Composition

Value	Count	Frequency (%)
United States	1027	68.1%
None Found	230	15.3%
India	29	< 0.1%
China	28	< 0.1%
Canada	28	< 0.1%
Switzerland	20	< 0.1%
United Kingdom	19	< 0.1%
France	14	< 0.1%
Australia	11	< 0.1%
Iceland	10	< 0.1%
Other values (32)	91	6.0%

informativeness

Numeric

Distinct count	9	**Mean**	5.8567
Unique (%)	< 0.1%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	8
Missing (n)	0	**Zeros (%)**	< 0.1%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	2
Q1	5
Median	6
Q3	7
95-th percentile	8
Maximum	8
Range	8
Interquartile range	2

Descriptive statistics

Standard deviation	1.778
Coef of variation	0.30359
Kurtosis	0.97813
Mean	5.8567
MAD	1.3697
Skewness	-1.1711
Sum	8826
Variance	3.1614
Memory size	11.9 KiB

monthsOld
Numeric

Distinct count	96	Mean	16.072
Unique (%)	6.4%	Minimum	1
Missing (%)	27.3%	Maximum	177
Missing (n)	411	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	1
5-th percentile	1
Q1	7
Median	11
Q3	12
95-th percentile	61
Maximum	177
Range	176
Interquartile range	5

Descriptive statistics

Standard deviation	22.214
Coef of variation	1.3822
Kurtosis	17.503
Mean	16.072
MAD	12.351
Skewness	3.8521
Sum	17615
Variance	493.48
Memory size	11.9 KiB

numericQuantifier
Numeric

Distinct count	1343	Mean	3.2032
Unique (%)	89.1%	Minimum	0
Missing (%)	0.0%	Maximum	17.134
Missing (n)	0	Zeros (%)	6.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	0
Q1	2.0211
Median	3.0817
Q3	4.2407
95-th percentile	6.2871
Maximum	17.134
Range	17.134
Interquartile range	2.2197

Descriptive statistics

Standard deviation	1.831
Coef of variation	0.57163
Kurtosis	3.1178
Mean	3.2032
MAD	1.4014
Skewness	0.82842
Sum	4827.2
Variance	3.3527
Memory size	11.9 KiB

region
Categorical

Distinct count	4
Unique (%)	< 0.1%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Frequency | Composition

Value	Count	Frequency (%)
North America	1058	70.2%
None Found	239	15.9%
Europe	107	7.1%
Asia/Oceania	103	6.8%

sentenceLength

Numeric

Distinct count	1463	**Mean**	26.937
Unique (%)	> 99.9%	**Minimum**	6.76
Missing (%)	0.0%	**Maximum**	62.682
Missing (n)	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	6.76
5-th percentile	19.358
Q1	23.497
Median	26.574
Q3	29.964
95-th percentile	35.245
Maximum	62.682
Range	55.922
Interquartile range	6.4679

Descriptive statistics

Standard deviation	5.2937
Coef of variation	0.19652
Kurtosis	4.1268
Mean	26.937
MAD	3.9498
Skewness	0.94234
Sum	40595
Variance	28.023
Memory size	11.9 KiB

TLD

Categorical

Distinct count	17
Unique (%)	< 0.1%
Missing (%)	< 0.1%
Missing (n)	1

Toggle details

Frequency | Composition

Value	Count	Frequency (%)
.com	1222	81.1%
.org	218	14.5%
.net	32	< 0.1%
.in	10	< 0.1%
.eu	5	< 0.1%
.co.uk	5	< 0.1%
.int	4	< 0.1%
.io	2	< 0.1%
.site	1	< 0.1%
.dhl	1	< 0.1%
Other values (6)	6	< 0.1%

vagueness
Numeric

Distinct count	1457	**Mean**	15.323
Unique (%)	> 99.9%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	39.164
Missing (n)	0	**Zeros (%)**	< 0.1%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	0
5-th percentile	7.3315
Q1	12.478
Median	15.393
Q3	18.416
95-th percentile	22.482
Maximum	39.164
Range	39.164
Interquartile range	5.9377

Descriptive statistics

Standard deviation	4.6216
Coef of variation	0.30162
Kurtosis	0.9958
Mean	15.323
MAD	3.5845
Skewness	-0.10199
Sum	23091
Variance	21.359
Memory size	11.9 KiB

wordCount
Numeric

Distinct count	1340	**Mean**	3282.4
Unique (%)	88.9%	**Minimum**	111
Missing (%)	0.0%	**Maximum**	23937
Missing (n)	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

Statistics | Histogram | Common Values | Extreme Values

Quantile statistics

Minimum	111
5-th percentile	429.6
Q1	1684
Median	2924
Q3	4369.5
95-th percentile	7300.3
Maximum	23937
Range	23826
Interquartile range	2685.5

Descriptive statistics

Standard deviation	2389.1
Coef of variation	0.72786
Kurtosis	11.478
Mean	3282.4
MAD	1698.9
Skewness	2.2841
Sum	4.9465e+06
Variance	5.7078e+06
Memory size	11.9 KiB

Appendix F

Python Files

Getting Dataset

August 9, 2019

```
In [1]: import pandas as pd
        import numpy as np
        from numpy import nan
        from datetime import datetime
        import matplotlib.pyplot as plt

        from langdetect import detect
        from difflib import SequenceMatcher
```

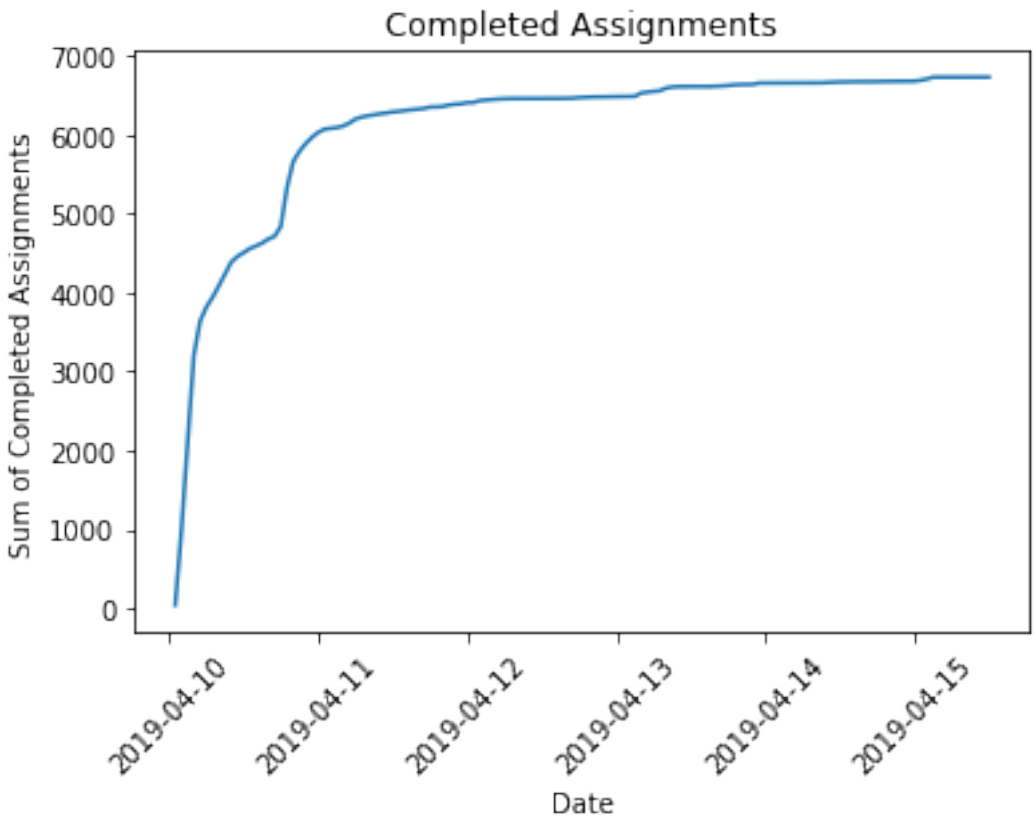
Import the data

```
In [17]: data = pd.read_csv('batchresults1604.csv')
```

Create Submission Proces graph

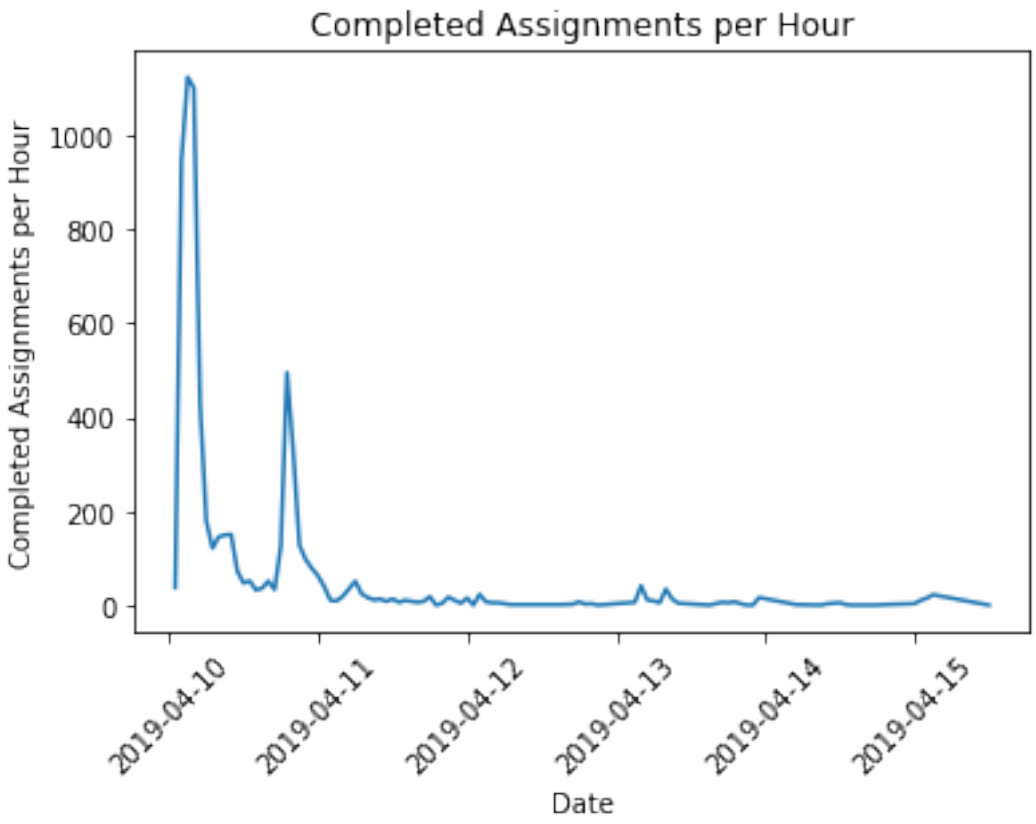
```
In [3]: def convert_time(x):
            return datetime.strptime(x[:13] + ' 2019', '%a %b %d %H %Y')

        data['counter'] = 1

        data['TimeGraph'] = data.SubmitTime.apply(lambda x: convert_time(x))

        plotdata = data[['TimeGraph','counter']].groupby(['TimeGraph'])['counter'].sum()
```

```
In [4]: plt.plot(plotdata.cumsum())
        plt.xticks(rotation=45)
        plt.xlabel('Date')
        plt.ylabel('Sum of Completed Assignments')
        plt.title('Completed Assignments')
        plt.show()
```



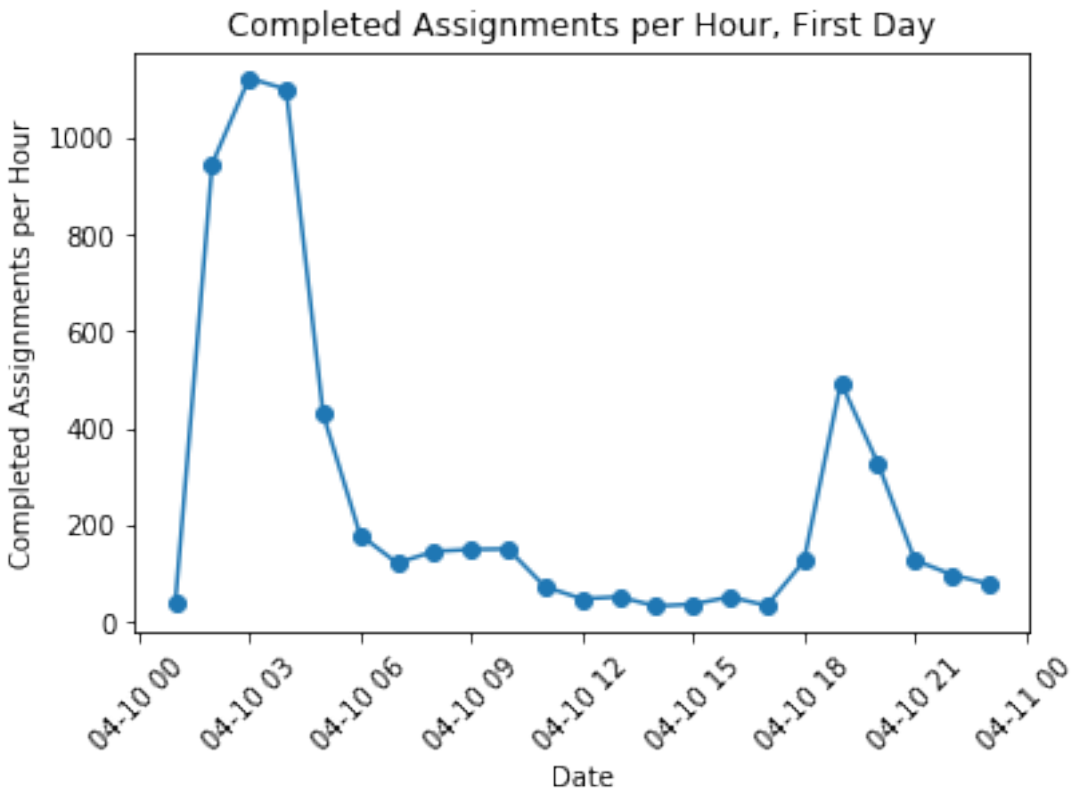
Sumission per hour graph

```
In [5]: plt.plot(plotdata)
        plt.xticks(rotation=45)
        plt.xlabel('Date')
        plt.ylabel('Completed Assignments per Hour')
        plt.title('Completed Assignments per Hour')
        plt.show()
```



Submission per hour, first day only

```
In [6]: plt.plot(plotdata[:23],'o-')
        plt.xticks(rotation=45)
        plt.xlabel('Date')
        plt.ylabel('Completed Assignments per Hour')
        plt.title('Completed Assignments per Hour, First Day')
        plt.show()
```



0.0.1 Creating the 2 subsets

First, preprocessing

```
In [18]: data['Answer.1.url'] = data['Answer.1.url'].str.lower()

         data = data[data.AssignmentStatus == 'Approved'].copy()

         data = data[data['Answer.3.full'].isna()==False]

         data['priv'] = data['Answer.3.full'].apply(lambda x: 'privacy' in x.lower())
         print('\n','Number of inputs without the word privacy:',len(data[data.priv==False]))
         data = data[data.priv == True]

         data['wordCount'] = data['Answer.3.full'].apply(lambda x: len(x.split()))
         print('\n','Number of too short inputs:',len(data[data.wordCount<11]))
         data = data[data.wordCount > 10]

         data['language'] = data['Answer.3.full'].apply(lambda x: detect(x))
         print('\n','Number of inputs in wrong language:',len(data[data.language!='en']))
         data = data[data.language == 'en']
```

```
        data.reset_index(inplace=True)
```

```
Number of inputs without the word privacy: 700

Number of too short inputs: 60

Number of inputs in wrong language: 9
```

Creating a column (count) which counts the amount of duplicates

```
In [19]: data['Check'] = data['Input.url'] + data['Answer.1.url'] + data['Answer.3.full']

         data['count'] = data.groupby('Check')['Check'].transform('count')
```

First subset, containing 135 privacy policies

```
In [20]: sub1 = data[data['count']>2][['Input.url','Answer.1.url','Answer.2.date',
                                      'Answer.3.full']].drop_duplicates('Answer.1.url')
         #sub1.to_csv('SampleData.csv',sep='|')

In [21]: len(sub1)

Out[21]: 127
```

Second subset, containing approx. 750 privacy policies.

```
In [22]: sub2 = data[data['count']>1][['Input.url','Answer.1.url','Answer.2.date',
                                     'Answer.3.full']].drop_duplicates('Answer.1.url')
         #sub2#.to_csv('SampleDataLarge.csv',sep='|')

In [23]: len(sub2)

Out[23]: 698
```

0.0.2 Creating the full dataset

First remove values from the subsets

```
In [24]: obtained = sub2['Input.url'].values

         dat = data[['Input.url','Answer.1.url', 'Answer.2.date', 'Answer.3.full']]
         dat = dat.sort_values('Input.url').rename(columns = {'Input.url':'input','Answer.1.url
         dat = dat[~dat.input.isin(obtained)]
```

Per url, select the best input to add to the final database

```
In [25]: dat['wordCount'] = dat.full.apply(lambda x: len(x.split()))

         def similar(a,b):
             return SequenceMatcher(None,a,b).ratio()

         fulldat = pd.DataFrame(columns = ['input','url','date','full'])

         count=0
         urls = dat.input.unique()
         #urls = ['aarp.org']

         for url in urls:
             count+=1
             copy = dat[dat.input == url].copy()
             copy.reset_index(inplace=True)

             if len(copy) == 3:
                 a = copy.iloc[0].wordCount
                 b = copy.iloc[1].wordCount
                 c = copy.iloc[2].wordCount

                 copy['lengthDifference'] = 0.01
                 copy.at[0,'lengthDifference'] = a/b
                 copy.at[1,'lengthDifference'] = b/c
                 copy.at[2,'lengthDifference'] = c/a

                 if any(x>=2 for x in copy.lengthDifference.values):
                     fulldat = fulldat.append(copy[copy.wordCount==copy.wordCount.max()].iloc[

                 else:
                     a = copy.full.iloc[0]
                     b = copy.full.iloc[1]
                     c = copy.full.iloc[2]

                     copy['similarity'] = 0.01
                     copy.at[0,'similarity'] = similar(a,b)
                     copy.at[1,'similarity'] = similar(b,c)
                     copy.at[2,'similarity'] = similar(c,a)

                     ind = copy[copy.similarity == copy.similarity.max()].index[0]

                     if pd.isna(copy.iloc[ind].date) == True:
                         if copy.date.isna().sum() < 3:
                             copy.at[ind,'date'] = copy.date[copy.date.isna()==False].values[0]

                     fulldat = fulldat.append(copy.iloc[ind,1:5])

             elif len(copy) == 2:
```

139

```
            ind = copy[copy.wordCount == copy.wordCount.max()].index[0]

            if pd.isna(copy.iloc[ind].date) == True:
                if copy.date.isna().sum() < 2:
                    copy.at[ind,'date'] == copy.date[copy.date.isna()==False].values[0]

            fulldat = fulldat.append(copy.iloc[ind,1:5])

        elif len(copy) == 1:
            fulldat = fulldat.append(copy.iloc[0,1:5])

        else:
            print('Error: found value with',len(copy),'entries.')

        print(count,'of',len(urls), 'policies scanned.', end='\r')
```

```
1131 of 1131 policies scanned.of 1131 policies scanned. of 1131 policies scanned.
```

Combine all entries (second subset and these newly added), remove all entries without 'privacy' in the url

```
In [62]: sub2 = sub2.rename(columns = {'Input.url':'input','Answer.1.url':'url','Answer.2.date

         full = pd.concat([sub2,fulldat]).drop_duplicates('url').drop_duplicates('input').drop
         full = full[full.url.isna() == False]

         full['check2'] = full.full.apply(lambda x: 'privacy policy' in x.lower())
         full['check3'] = full.url.apply(lambda x: 'privacy' in x)

         full = full.sort_values('input').reset_index(drop=True)
         nostate = full[(full.check2 == False)&(full.check3==False)].index.values
         full = full.drop(nostate)
         final = full.sort_values('input').reset_index(drop=True)
```

After manual assessment of all links that did not contain the word privacy, a number of entries are manually removed. This was done manually because approx. 75% of the entries were still valid, and can be kept using this method.

```
In [63]: final.iloc[:,:4].to_csv('FullDataset.csv', sep = '|')
```

NLP Variables

August 9, 2019

The code below uses the dataset of 1510 privacy statements, which is retrieved after cleaning the full MTurk dataset, and extracts all variables used in the research.

1 Start-up

```
In [1]: import pandas as pd
        import numpy as np
        from numpy import nan as NaN
        import random
        import matplotlib.pyplot as plt
        from collections import Counter
        import seaborn as sns
        from readability import Readability
        from country_list import countries_for_language
        import datetime
        import pyap

        from nltk.tokenize import sent_tokenize, word_tokenize

        import gensim

        import spacy
        from spacy import displacy

        import en_core_web_md
        nlp = en_core_web_md.load()

        from spacy.tokenizer import Tokenizer
        tokenizer = Tokenizer(nlp.vocab)
```

Importing and combining datasets

```
In [2]: samp = pd.read_csv('FullDataset.csv', sep = '|').iloc[:,1:]
        fullcat = pd.read_csv('0705FullCategories.csv')

        samp = samp.merge(fullcat[['input','cat','score']], on = 'input')
```

```
In [3]: def dateconverter(x):
            if type(x) == str:
                if x[2]=='-':
                    return np.datetime64(x[6:10]+x[2:5]+x[5]+x[:2])
                else:
                    return np.datetime64(x)
            else:
                return x

        samp.date = samp.date.apply(lambda x: dateconverter(x))
```

```
In [4]: def months_ago(x):
            if pd.isnull(x):
                return x
            else:
                date = (np.datetime64('2019-04') - np.datetime64(x,'M')) / np.timedelta64(1,'M
                if date > 0:
                    return date
                else:
                    return np.nan

        samp['monthsOld'] = samp.date.apply(lambda x: months_ago(x))

        for i in samp[samp.monthsOld > 200].index:
            samp.at[i,'monthsOld'] = np.nan
```

```
In [5]: def return_tld(x):
            for tld in tlds:
                if tld in x:
                    return tld

        tlds = ['.com','.net','.org','.eu','.int','.in','.co.uk','.site','.sg','.dhl','.edu',
                '.kpmg','.us','.co.jp','.io','.nl']

        samp['TLD'] = samp.url.apply(lambda x: return_tld(x))
```

2 Simple methods

```
In [6]: #vague terms from A Theory of Vagueness and Privacy Risk Perception (Bhatia et. al, 20
        VT = ['generally','mostly','widely','general','commonly',
              'usually','normally','typically','largely','often',
              'may','might','can','could','would','likely','possible',
              'possibly']

        #conditionality
        C = ['depending','necessary','appropriate','inappropriate','as needed']
        #generalization
        G = ['generally','mostly','widely','general','commonly','usually','normally','typicall
```

```
#modality
M = ['may','might','can','could','would','likely','possible','possibly']
#numeric quantifier
N = ['certain','some','most']


#third party
tp = ['third party','third parties','third-party','third-parties','business partners']

#opt in/out combinations
opt =['opt in','opt out','opt-in','opt-out']
opt2 = [['your','choice'],['you','choice'],['you','choose']]

#special audiences
specaud = ['children','under 13','california privacy rights','coppa']

#policy changes
policy = ['policy','statement','notice']
change = ['change','changes','update','revise']

#data retention
ret = [['data','retention'],['information','retention'],['long','keep','data'],['long'
       ['retain','data'],['retain','information']]

#access rights
ar = [['right','access'],['right','request'],['request','access']]

#consent
cons = [['legal','bases'],['legal','basis'],['lawful','bases'],['lawful','basis']]

#external organisations
social = ['social media','social network']
nai = ['nai','network advertising initiative']
daa = ['daa', 'digital advertising alliance']

#words used when information is shared
shareWords = ['share','disclose']

dpo = ['dpo', 'data protection officer']
datcon = ['data controller']
datpros = ['data processor']
gdpr = ['gdpr', 'general data protection regulation']



newdata=[]
```

```
for pp in range(len(samp)): #for each privacy policy
    #define the text
    text = samp.full[pp].lower()
    words = text.split() #cleaned words
    sentences = sent_tokenize(text)

    #test a number of readability scores
    r = Readability(text)
    if len(words) > 100:
        score = r.flesch().score
        ease = r.flesch().ease
    else:
        score = NaN
        ease = NaN

    if pd.isnull(samp.monthsOld[pp]):
        aftGDPR = False
    else:
        aftGDPR = samp.monthsOld[pp] < 16

    #count the number of vague terms
    count = sum(w in VT for w in words)
    countC = sum(w in C for w in words)
    countG = sum(w in G for w in words)
    countM = sum(w in M for w in words)
    countN = sum(w in N for w in words)


    #check coverage
    arv = False
    changev = False
    retainv = False

    publicInterest = False
    unambiguousConsent = False


    if 'public obligations' in text:
        publicInterest = True

    for s in sentences:

        if any(x in s for x in policy):
            if any(x in s for x in change):
                changev = True

        if any(all(x in s for x in comb) for comb in ret):
            retainv = True      144
```

```
        if any(all(x in s for x in comb) for comb in ar):
            arv = True

        for w in nlp(s):
            if w.lemma_ in shareWords:
                for w in nlp(s):
                    if w.lemma_ == 'authority':
                        publicInterest = True

        if any(all(x in s for x in comb) for comb in cons):
            if 'consent' in s:
                unambiguousConsent = True

    #combine all data
    newdata.append([samp.url[pp],
                    len(words),
                    len(words)/len(sentences),
                    score,
                    ease,
                    (count/len(words))*1000,
                    (countC/len(words))*1000,
                    (countG/len(words))*1000,
                    (countM/len(words))*1000,
                    (countN/len(words))*1000,
                    'vital interest' in text,
                    'legitimate interest' in text,
                    publicInterest,
                    'contract ' in text,
                   'legal obligations' in text,
                    unambiguousConsent,
                    arv,
                    any(x in text for x in dpo),
                    any(x in text for x in datcon),
                    any(x in text for x in datpros),
                    aftGDPR,
                    any(x in text for x in tp),
                    any(x in text for x in opt) or any(all(x in s for x in comb) for c
                    'security' in text,
                    any(x in text for x in specaud),
                    'privacy shield' in text,
                    retainv,
                    changev,
                    'google' in text,
                     'facebook' in text,
                     'amazon' in text,
                     'microsoft' in text,
                    'criteo' in text,
```

```
                                'adobe' in text,
                                'doubleclick' in text,
                                'nielsen' in text,
                                'paypal' in text,
                                'apple' in text,
                                'stripe' in text,
                                'yahoo' in text,
                                any(x in text for x in social),
                                any(x in text for x in nai),
                                any(x in text for x in daa),
                                'android' in text,
                               any('@' in word and '.' in word for word in text.split())

                              ])

            print(pp+1, 'of',len(samp),"policies scanned.", end='\r')

        #put it all in a Dataframe
        ppdata = pd.DataFrame(newdata,columns=['site','wordCount','sentenceLength','fleschScor
                                               'vagueness','conditionality','generalization','
                                               'vitalInterest','legitimateInterest','publicInt
                                               'contractualNecessity', 'legalObligations','una
                                               'dpo','dataController','dataProcessor','afterGD
                                               'thirdParty','choices','security','specificAudi
                                                'privacyShield','dataRetention','policyChange'
                                               'google','facebook','amazon','microsoft',
                                               'criteo','adobe','doubleClick','nielsen','paypa
                                               'socialMedia','nai','daa','android','emailProvi
```

```
1518 of 1518 policies scanned.
```

3 Complex methods

```
In [14]: colWords = ['collect','store','gather','keep','you provide','log','ask','request',
                     'receive','identify','derive','record','include','give']
         shareWords = ['share','disclose']

         def remove_punct(x):
             return ' '.join(str(w) for w in nlp(x) if not w.is_punct)

         col = samp[['url','full','TLD','cat']].rename(columns={'url':'site'})

         dataWords = ['information','data']

         compInfoWords = ['operating system','browser type','system type','unique device','bro
                          'ip address','internet protocol','device information','ip-address']
         contactInfoWords = ['contact information',' name','email address','home address','pos
```

```
                    ,'telephone number']
webTrackingWords = [' logs ','web beacons','tracking cookies','device fingerprinting'
                    'clear gif']
personalInfoWords = ['personal information','personally identifiable information','pi
thirdPartyWords = ['third party','third parties','third-party','third-parties','busin
financialWords = ['financial information','credit card','billing information','debit
demographyWords = ['demographic information',' gender',' age','education','profession
                    'marital status']

notifyWords = ['notice','notify']
emailWords = ['e-mail address','e-mail','email','e mail']

USA = ['U.S.','US','USA']

col['computerInformation'] = False
col['contactInformation'] = False
col['webTracking'] = False
col['personalInformation'] = False
col['dataCombining'] = False
col['financialInformation'] = False
col['location'] = False
col['demographicInformation'] = False

col['computerInformationShare'] = False
col['contactInformationShare'] = False
col['webTrackingShare'] = False
col['personalInformationShare'] = False
col['financialInformationShare'] = False
col['locationShare'] = False
col['demographicInformationShare'] = False
col['generalShare'] = False
col['changeNotification'] = False
col['headquarters'] = 'None Found'
col['addressProvided'] = False

countries = list(x[1] for x in countries_for_language('en'))
countries.append('U.S.')
countries.append('US')
countries.append('USA')

states = [" AK ", " AL ", " AR ", " AZ ", " CA ", " CO ", " CT ", " DE ", " FL ", " G
          " IL ", " IN ", " KS ", " KY ", " LA ", " MA ", " MD ", " ME ", " MI ", " M
          " NC ", " ND ", " NE ", " NH ", " NJ ", " NM ", " NV ", " NY ", " OH ", " O
          " SC ", " SD ", " TN ", " TX ", " UT ", " VA ", " VT ", " WA ", " WI ", " W

for i in col.index:
    print(i+1, 'of',len(col),"policies scanned.", end='\r')
```

```
PS = col.iloc[i].full


locationCheck = False
compInfoCheck = False
contactInfoCheck = False
webTrackingCheck = False
personalInfoCheck = False
financialCheck = False
combineCheck = False
demographyCheck = False

locationCheckS = False
compInfoCheckS = False
contactInfoCheckS = False
webTrackingCheckS = False
personalInfoCheckS = False
financialCheckS = False
demographyCheckS = False
generalS = False
notifyCheck = False
addressProvided = False
stateCheck = False

notifyVar = 0

countrylist = []


for sent in sent_tokenize(PS):


    sentNP = remove_punct(sent.lower())

    if not locationCheck & locationCheckS:
        if 'location' in sentNP:
            if any(token.lemma_ in colWords for token in nlp(sent) if not locatio
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    locationCheck = True
            if any(token.lemma_ in shareWords for token in nlp(sent) if not locat
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    locationCheckS = True

    if not compInfoCheck & compInfoCheckS:
        if any(x in sentNP for x in compInfoWords):
            if any(token.lemma_ in colWords for token in nlp(sent) if not compInf
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    compInfoCheck = True
```

```
            if any(token.lemma_ in shareWords for token in nlp(sent) if not compI
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    compInfoCheckS = True

    if not contactInfoCheck & contactInfoCheckS:
        if any(x in sentNP for x in contactInfoWords):
            if any(token.lemma_ in colWords for token in nlp(sent) if not contact
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    contactInfoCheck = True
            if any(token.lemma_ in shareWords for token in nlp(sent) if not conta
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    contactInfoCheckS = True

    if not webTrackingCheck & webTrackingCheckS:
        if any(x in sentNP for x in webTrackingWords):
            if any(token.lemma_ in colWords for token in nlp(sent) if not webTrac
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    webTrackingCheck = True
            if any(token.lemma_ in shareWords for token in nlp(sent) if not webTra
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    webTrackingCheckS = True

    if not personalInfoCheck & personalInfoCheckS:
        if any(x in sentNP for x in personalInfoWords):
            if any(token.lemma_ in colWords for token in nlp(sent) if not personal
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0 and 'no
                    personalInfoCheck = True
            if any(token.lemma_ in shareWords for token in nlp(sent) if not perso
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0 and 'no
                    personalInfoCheckS = True

    if not financialCheck & financialCheckS:
        if any(x in sentNP for x in financialWords):
            if any(token.lemma_ in colWords for token in nlp(sent) if not financia
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    financialCheck = True
            if any(token.lemma_ in shareWords for token in nlp(sent) if not finan
                if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                    financialCheckS = True


    if not combineCheck:
        if 'combine' in sentNP:
            if any(x in sentNP for x in thirdPartyWords):
                if any(token.lemma_ in colWords for token in nlp(sent) if not com
                    if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                        combineCheck = True
```

```
        if not demographyCheck & demographyCheckS:
            if any(x in sentNP for x in demographyWords):
                if any(token.lemma_ in colWords for token in nlp(sent) if not demograp
                    if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                        demographyCheck = True
                if any(token.lemma_ in shareWords for token in nlp(sent) if not demogr
                    if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                        demographyCheckS = True

        if not generalS:
            if any(x in sentNP for x in dataWords):
                if any(token.lemma_ in shareWords for token in nlp(sentNP) if not gene
                    if sum([token.dep_ == 'neg' for token in nlp(sent)]) == 0:
                        generalS = True

        if not notifyCheck:
            if notifyVar == 0:
                if any(x in sentNP for x in policy):
                    if any(x in sentNP for x in change):
                        notifyVar = 1
            if notifyVar != 0 and notifyVar < 7:
                if any(x in sentNP for x in notifyWords):
                    if any(x in sentNP for x in emailWords):
                        notifyCheck = True
                notifyVar += 1
            else:
                notifyVar = 0

        for x in countries:
            if x in sent:
                countrylist.append(x)
                for x in sent.split():
                    if x.isdigit():
                        addressProvided = True

        for x in states:
            if x in sent:
                for word in sent.split():
                    if word.isdigit():
                        stateCheck = True
                        addressProvided = True

    col.at[i,'location'] = locationCheck
    col.at[i,'computerInformation'] = compInfoCheck
    col.at[i,'contactInformation'] = contactInfoCheck
    col.at[i,'webTracking'] = webTrackingCheck
    col.at[i,'personalInformation'] = personalInfoCheck
    col.at[i,'financialInformation'] = financialCheck
```

```
                col.at[i,'dataCombining'] = combineCheck
                col.at[i,'demographicInformation'] = demographyCheck

                col.at[i,'locationShare'] = locationCheckS
                col.at[i,'computerInformationShare'] = compInfoCheckS
                col.at[i,'contactInformationShare'] = contactInfoCheckS
                col.at[i,'webTrackingShare'] = webTrackingCheckS
                col.at[i,'personalInformationShare'] = personalInfoCheckS
                col.at[i,'financialInformationShare'] = financialCheckS
                col.at[i,'demographicInformationShare'] = demographyCheckS
                col.at[i,'generalShare'] = generalS
                col.at[i,'changeNotification'] = notifyCheck

                c=0
                countrylistFix = []
                for x in countrylist:
                    if x not in USA:
                        countrylistFix.append(x)
                    else:
                        c += 1
                fix = ['United States'] * c
                countrylistFix.extend(fix)
                countrylist = countrylistFix

                if stateCheck:
                    col.at[i,'headquarters'] = 'United States'
                if len(countrylist) > 0:
                    topCountry = max(set(countrylist), key=countrylist.count)
                    col.at[i,'headquarters'] = topCountry

                if addressProvided:
                    col.at[i,'addressProvided'] = addressProvided
                else:
                    if topCountry == 'United States':
                        col.at[i,'addressProvided'] = len(pyap.parse(samp.full.iloc[i],country =
                    elif topCountry == 'Canada':
                        col.at[i,'addressProvided'] = len(pyap.parse(samp.full.iloc[i],country =
                    else:
                        col.at[i,'addressProvided'] = addressProvided
```

```
1518 of 1518 policies scanned.
```

4 Merge all data

```
In [15]: final = ppdata.merge(col.set_index('site').iloc[:,1:] , on = 'site')
```

5 Add final values, based on retrieved data

Firstly, age of the privacy statements is added to the dataset. Next, informativeness, data colletion and data sharing are added as combined variables.

```
In [16]: final['monthsOld'] = samp.monthsOld

         final['informativeness'] = final.iloc[:,np.r_[21:28,62]].sum(axis=1)
         final['dataCollection'] = final.iloc[:,47:55].sum(axis=1)
         final['dataSharing'] = final.iloc[:,55:63].sum(axis=1)
```

Combining the headquarter countries into three regions.

```
In [17]: NA = ['United States','Jersey','Canada','Mexico','Cayman Islands']
         EU = ['Switzerland','United Kingdom','France','Italy','Iceland','Germany','Finland','
               'Lithuania','Spain','Isle of Man','Poland','Sweden','Cyprus','Malta','Ireland',
         AS = ['Taiwan','China','Qatar','India','Singapore','Russia','Thailand','Australia','U
              'Japan','Israel','New Zealand']

         def define_region(x):
             if x in NA:
                 return 'North America'
             elif x in EU:
                 return 'Europe'
             elif x in AS:
                 return 'Asia/Oceania'
             else:
                 return 'None Found'

         final['region'] = final['headquarters'].apply(lambda x: define_region(x))
```

Exporting the output to csv.

```
In [19]: final.to_csv('FullDataOutput.csv', sep = '|')
```

5.1 Code used for manual validation

Code below copies ten random statements, these are manually verified in excel.

```
In [ ]: validationSet = set()
        while len(validationSet) < 10:
            validationSet.add(np.random.randint(150))

        final.iloc[list(validationSet)][['site','emailProvided']].to_clipboard()
```

From hereon further, the original code contained a lot of testing code, te assess why certain privacy statements produced false postives or false negatives. This code has been removed as it does not directly contribute to the data.