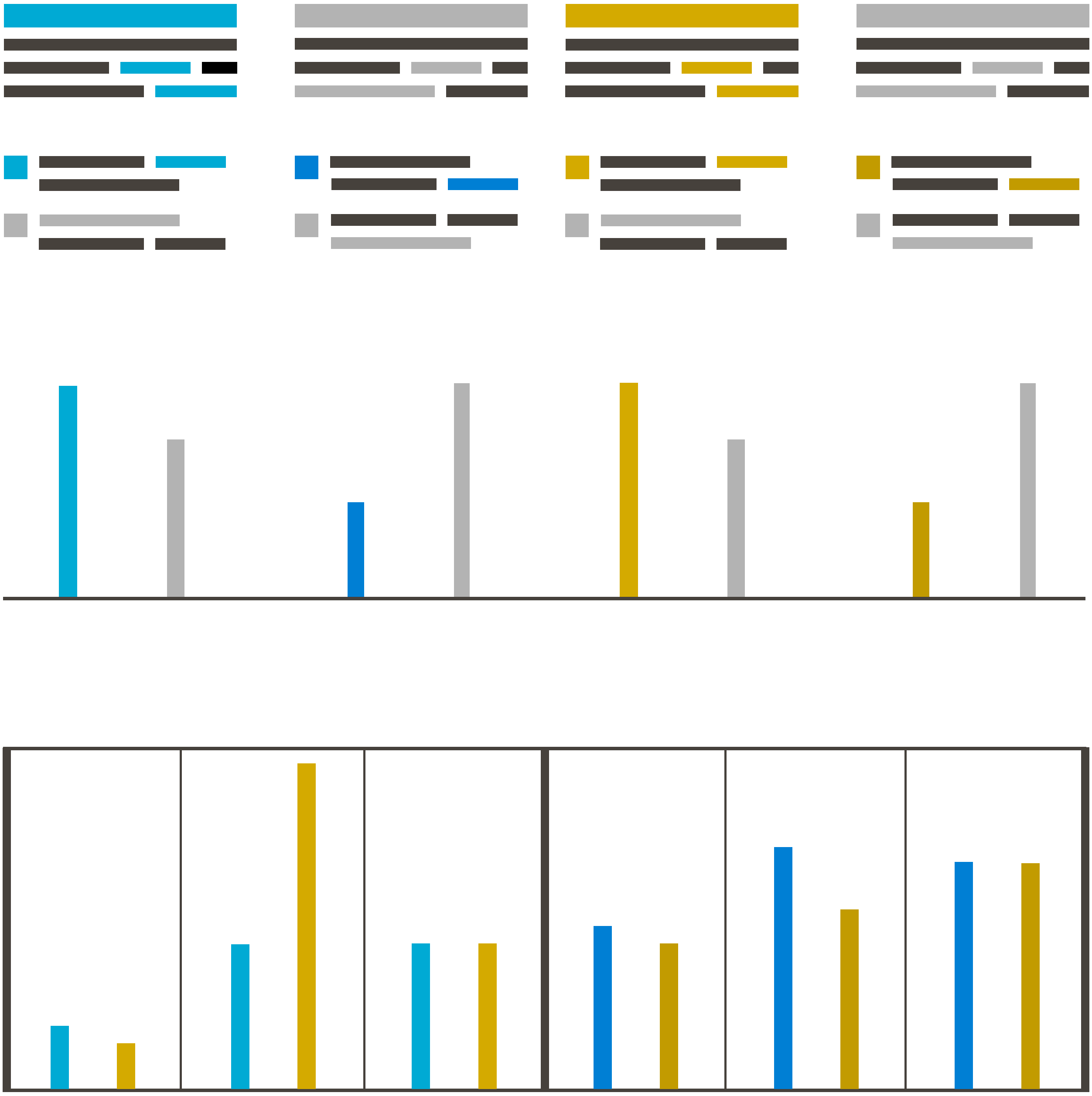
TOPIC MODELLING IN ONLINE DISCUSSIONS

Analysis of the Developments within the Dutch Privacy Debate on News Websites



Chris Emmery

TOPIC MODELLING IN ONLINE DISCUSSIONS

ANALYSIS OF THE DEVELOPMENTS WITHIN THE DUTCH PRIVACY DEBATE ON NEWS WEBSITES

Chris Emmery

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COMMUNICATION AND INFORMATION SCIENCES,
MASTER TRACK HUMAN ASPECTS OF INFORMATION TECHNOLOGY,
AT THE SCHOOL OF HUMANITIES
OF TILBURG UNIVERSITY

School of Humanities
Department of Communication and Information Sciences
Tilburg Center for Cognition and Communication (TiCC)
Tilburg University

ANR:
609661

SUPERVISOR:
Dr. M.M. van Zaanen

SECOND READER:
Prof. dr. E.O. Postma

LOCATION:
Tilburg

DATE:
August 22, 2014

This thesis was typesetted in LaTeX based on the "classically styled" thesis package `classicthesis` by André Miede. It was written using predominantly (with exception of only a few books and papers) open-source information and software.

"What happens to you here is for ever."

– George Orwell, *1984*

ABSTRACT

The current research focuses on the computational analysis of online discussions. Specifically, it is investigated how the disclosures surrounding the NSA's mass-surveillance activities impacted the Dutch privacy debate on online news communities. The main research aim is to create a better understanding of the effectiveness of topic modelling in online discussions. To achieve this, data is collected from Nu.nl and Tweakers.net, resulting in a dataset containing news articles and associated comments for both sources. Given the distinct news scope of these sources, their users are argued to differ in terms of topic experience. Accordingly, Tweakers.net was argued to be the experienced group, and Nu.nl non-experienced. It is hypothesized that topic exposure through news surrounding the NSA increased overall privacy discussion, and influenced both of these groups differently as a result of their varying topic experience.

Labelled LDA was trained to infer topic labels for these sets, allowing a comparison between the sources their discussion frequency analysis over time. Exposure to the topic of privacy was shown to directly affect related discussion frequency for both communities. This was most notable in articles on the topic of privacy for Nu.nl. For Tweakers.net, the effect was strongest in the spread of discussion to non-related articles. The effect equalized, and was shown to be persistent, over time for both sources. Low topic awareness is therefore argued to be linked to a strong reaction on the initial news, and a delayed discussion spread.

Keywords:
supervised topic modelling, online discussion, LDA, digital privacy

PREFACE

This document condenses most of the developments I myself, and my academic interest in particular, have gone through over the course of three years. Ever since high school I have had trouble finding the challenges and intrigue I did find far more often in my hobbies. Neither subjects nor teachers succeeded in sparking my interest in their material over the many video games I played, and my general interest in technology subsequently did not find its way. Even during my Bachelor I still had strong doubts if this was ever going to happen. Then along came the third year of my Bachelor, which felt as though I was finally handed the hammer and shown my way to the anvil.

In three years, I discovered the beauty and joy of programming, and have come to admire data analysis in broad spectrum. I became increasingly fanatic about Linux and open-source. Each lecture I became more and more convinced that I found my place; I thoroughly enjoyed writing both my theses and applying theoretical knowledge, as well as practical skills I gained over the years. I participated in research groups, extracurricular activities, and started doing course-related and development work for fun. Most importantly, I had the honour to meet three years worth of amazing people that formed the Human Aspects of Information Technology students, as well as my supervisors, who have had a key role in stimulating me to prolong my work in academia. At least, for as long as I have the opportunity.

Hence, writing this thesis and finishing my studies with such motivation and enjoyment would not have been possible if it was not for them. In the year leading up to his retirement, Hans Paijmans was, despite voicing his discontent with this difficult task on many occasions, my Bachelor supervisor. Paai was the first one to show me Linux, learn me to program in AWK and to show me the combination of humanities-related research and programming. His teachings, patience and confidence in my abilities as a researcher have proven to be that hammer I needed. During my Master, I have learned an unthinkable amount from both Menno van Zaanen and Eric Postma's lectures. Though both have helped me a great deal in the process of my thesis and its aftermath, Menno in particular has gone above and beyond as a supervisor. The many (IRC) chats we had regarding the thesis, the opportunities I was given to become a student assistant, and being included in research groups, have proven invaluable. The three of you have my deepest, and more than well deserved gratitude for making this all turn out the way it did.

It should go without saying that many, many thanks go out to my family and friends, and especially to my lovely girlfriend Monique for supporting me in the important decisions during my studies, and in the periods where I was more interwoven with my computer than anything else. Finally, nearly crucial were the (almost) three years of more than fulltime occupation of the HAIT lab I shared in varying combinations with Vincent, Nanne, Rick, Suzanne and Ákos. Apart from the many study and thesis related discussions, tech rants, bar visits, whisky nights, and game sessions we had, you have most of all proven very dear friends, and I am grateful to be the last one to close the lab door from a range of such awesome students.

またね

CONTENTS

"Now, we must all fear evil men. But there is another kind of evil which we must fear most, and that is the indifference of good men."

– Monsignor, *Boondock Saints*

1 INTRODUCTION

"We live in a post-Snowden era". It is a description of the crossroad our information society has been placed on, frequently used by Jacob Appelbaum; computer security researcher, hacker, and core member of the TOR project[1]. Between May and June 2013, Edward Snowden, a system administrator and former employee of the CIA, and most notably contractor for the NSA[2], leaked classified documents regarding global surveillance activities of intelligence agencies in primarily the United States and Britain. Amongst these, programs such as PRISM, XKeyscore, Tempora, MUSCULAR, and FASCIA were revealed (Appelbaum, 2013; Der Spiegel, 2013; Etzioni, 2014; McGowan, 2014; The Guardian, 2013). All of them amass to a single goal: the ability to collect, store, and interpret any piece of digital personal information.

Mass surveillance, as depicted in fictional works such as Orwell (1949) and Stephenson (2011), could be seen as optimistic in comparison to the reported activities of the NSA and GHCQ[3]. With the possibility of any internet-connected device being under surveillance by dragnet techniques, digital privacy can quite reasonably be declared void for the general public. Browser activity, search queries, social media information and messages, phone calls, microphone audio, monitor as well as camera and webcam images, location data; any step taken while being connected to the Internet seems prone to be recorded, stored and analysed if it to some degree would benefit the security of countries that have the means and intention to do so.

Since the still ongoing releases, an important public debate has sparked in America; it primarily concerns whether or not the NSA's actions are justified, and could very well determine which direction global information society will take at the aforementioned crossroad: the path of opting for public tolerance of widespread surveillance solely in the name of the war on terror, or a transparent intelligence service for legitimate specific targets. Hence, a strikingly contrasting

[1] www.torproject.org

[2] National Security Agency (USA).

[3] Government Communications Headquarters (UK).

observation is that the Dutch public debate[4] has ostensibly been absent. More so due to the fact that media attention to this topic, on the other hand, has been persistent. Even before Snowden, a sixty minute documentary on the surfacing Dutch surveillance state and its implications, named Panopticon (Vlemmix, 2012), was released and became the year's best rated and most viewed online documentary. Along with the extensive coverage surrounding the stream of leaked documents by newspapers, as well as critical responses from several civil rights movements, and even members of political parties, this does however provide evidence for the subject being spread enough for both offline and online discussion[5] to have manifested.

This thesis will tie into these events by aiming towards a method of identification and quantitative analysis of the associated public discussion; more specifically with the focus on the development of this discussion over time. It is hypothesized that the disclosures surrounding Edward Snowden, and the increasing concern regarding the activities of both the NSA and GHCQ, has increased the discussion concerning the topic of privacy, and digital privacy in particular. While some have both qualitatively as quantitatively analysed public discussion regarding a certain topic over time (Wang, Wang, & Zhu, 2013), others have sought to analyse the sentiment in discussion to extract the public opinion (Kaur & Gupta, 2013), in for example consumer confidence and political opinions (O'Connor & Balasubramanyan, 2010). As to our knowledge, the research in this thesis touches upon novel ground by assessing how discussion evolves in the period of more than a year, and more specifically by trying to quantify the changes in topic exposure and correlation with the frequency of public discussion. By opting for an analysis of online news sources that are accompanied with a platform of discussion, models to identify and classify these discussion streams are also applied to a source that has not yet been used in previous research. In this sense, it is hypothesized that the field of Topic Modelling can be of use as a framework to effectively assess developments in public discussion and topic exposure through statistical analysis.

To facilitate a thorough understanding of the multidisciplinary topic that this thesis covers, the background will be divided in two chapters. In Chapter 2, a concise historic overview of online data collection is given, after which a resulting shift in the principle of privacy is related to the current state of Dutch privacy. From here, the research questions will be given, and the approach to tackle the challenges they hold will be discussed through several observations in

[4] Here, public debate is defined as ongoing public discussion between the public, or public organizations and the government, as well as the topic being an active point on the House of Representatives its debate agenda and therefore not solely restricted to responses on individual incidents.

[5] Note that this form of discussion can be considered 'behind closed doors' and is therefore not in the scope of public debate.

the computational field. In Chapter 3, the computational methods extracted from Information Retrieval, as well as Machine Learning will be thoroughly discussed. After an introduction to the field of Topic Modelling is given in Section 3.2, Latent Dirichlet Allocation will be discussed in 3.3. Labelled LDA in particular will be used as Topic Model for inferring labels onto the collected data, for which the procedure will be discussed in Chapter 4. The requirements and collection process of the dataset will be handled in Section 4.1, whereas the steps taken preparation for the inference process are the topic of Section 4.2. Hereafter, the results of the two conducted experiments will be reported and discussed consecutively: the Model Evaluation in Chapter 5 and the Analysis in Chapter 6. Finally, the thesis is concluded in Chapter 7.

Dialectic13 – "They're so proud of themselves. "Exposing" all the dirty little secrets of the U.S. government and the likes of STRATFOR, et al. Brilliant. Why don't they just send correspondence straight to Moscow and Beijing while they're at it. Shortsighted fools."

Overwelmdn underated – "they aren't even enemies of the US, stop playing Goldeneye lol"

– The YouTube Comment Section

2 THE SNOWDEN DEBATE

Edward Snowden's actions allow for a novel view on discussion surrounding intelligence services, sharing information on the 'free' Internet and even more importantly the concept of digital privacy. However, in order to assess the Snowden debate, it is important to review the events that have led up to the spark of this widespread public debate, as well as the key issues that have delayed this discussion for a long time. Therefore, Section 2.1 will give a compact overview of relevant developments that the Internet brought, and how those moved away from the initial view and principles behind the Internet. From here, it will become apparent that the principle of privacy, the primary topic of the Snowden debate, has shifted. Section 2.2, accordingly, deals with the societal interwoven concept of privacy and its classical definition having decayed through the rise of data collection. Herein, it will also be argued that the lines between public and private space have blurred, and laws and jurisdiction upholding the rights to privacy seem incapable of providing a framework for digital privacy. This theory will help to frame the state of Dutch privacy regulations and the related discussion amongst the country's citizens in Section 2.2.4. Assessing these will then lead to the research question in Section 2.3.

> *"A strange game. The only winning move is not to play."*
>
> – WOPR, WarGames

2.1 RISE OF AN INFORMATION SOCIETY

The history of the Internet spans many different fields and pieces of history, well recorded and overviewed in various other sources (e.g. Leiner, Cerf, & Clark, 2009), from which only the relevant parts will be discussed here. Hence, the focus in this section lies partly on the core principles, and primarily on the history of online data collection and mining. The overview is aimed at providing an intuition regarding the gradual increase in use of personal information in services provided on the Internet, and its integration in offline tasks. As will become clear, the free Internet's inherent ideals have, through a partial neglect of these, ultimately led to their own undoing.

2.1.1 *Principles of the Internet*

Springing out of governmentally financed research project from 1962 onwards, the initial goal of the Internet was to simplify communication and information sharing. The army had adapted these features in a way that communication was secure and separated by 1980. When it opened up to the public in 1993[1], the sense of an independent utopia of information where everything was possible came into existence (Barlow, 1996). However, the Internet turned out both economically and politically important, and the defiance of national borders posed many issues. This resulted in the almost anarchistic freedom being limited through Internet Governance (Uerpmann-Wittzack, 2010) and Internet Law, which yielded objective principles ought to be upheld by law, from which three are most relevant to the Snowden debate.

First and foremost, the Internet is supposed to be open, accessible for anyone with a connection, its content unfiltered and globally reachable. This is the concept of Net Neutrality as described by the FCC (2010)[2], imposed in many countries through various platforms. Secondly, the Principle of Internet Freedom allows in essence the freedom of communication and expression, as a universal right. Finally, the Principle of Privacy states that emails or other data which is transmitted by, or accessible through the Internet are private, unless destined for public access. The latter two are part of Articles written by the European Court of Human Rights.

[1] The United States opened up the Internet to companies and individuals. XS4all did the same in the Netherlands.

[2] In contrast, at the time of writing, the FCC are trying to weaken Net Neutrality in the United States through implementing a higher cost 'fast lane' (FCC, 2014).

It could be argued that judging from these principles, the initial ideal behind the Internet is still being upheld; internet users are free and anonymous if they so desire. However, these documents also teach us that these principles have some leeway: "*Unlike rules, principles do not require strict observance. Due to their broad scope, they easily collide with other principles or interests. In this case, the principle has to be realized as far as this is possible under the given legal and factual circumstances.*" (Uerpmann-Wittzack, 2010). Conclusively, when these principles interfere with rights of others, national security, public order and morals, they are reassessed. Logically, and as we now know, governmental agencies utilize these jurisdictional loopholes such as the PATRIOT act[3] for the NSA. However, the collection and analysis of data by these agencies would not have been able to be performed at this scale, if it was not for both corporations neglecting many of the above mentioned principles, and the internet users themselves agreeing and cooperating with the developments in the field of corporate data collection and mining. To understand the gradual process that led to the undoing of the users their own rights, we must trace back to the beginning of the Internet.

2.1.2 *Data Collection*

In the very beginning of the Internet opening up to consumers, services such as Yahoo and AltaVista aimed to provide 'yellow pages' of the existing web content. At that time, this content was made up of companies or individuals setting up network connected pages to provide and share information. Effectively indexing these pages then became an important goal for many start-ups, from which Google evidently turned out to have the most effective method (Page, Brin, Motwani, & Winograd, 1999) and resources to do so. In order to grow in scale, companies similar to Google, who only offered services online, had to devise some way to keep scaling to their increasing popularity and stay that popular. With improving functionality came many techniques which we now know from the Information Search domain (Manning, Raghavan, & Schütze, 2008), whereas the funding for their services was provided through selling and serving users advertisements relevant to their search results (Fain & Pedersen, 2006).

With growing internet connections and cheaper storage possibilities came an important concept that encompassed features of the websites created around 2004 and onwards: Web 2.0. Allowing anyone with internet access to share, interact and collaborate, the online information stream exploded through blogs, wikis, social networking sites, as well as video and web hosting services. This interaction initially gave any user a medium without borders to share thoughts and opinions, anonymously, associated with a nickname only. The

[3] http://www.gpo.gov/fdsys/pkg/PLAW-107publ56/pdf/PLAW-107publ56.pdf

Internet became a source with 'unlimited' amounts of free information which, following this ideal, gave rise to the first digital pirates that shared copyrighted material for free (Liebowitz, 2012). Online services began to develop new methods of automatically interpreting the large sets of new user-generated data to improve their functionalities and performance, for which many techniques from the field of Data Mining were employed.

2.1.3 *Data Mining*

In time, many of these websites started connecting your person to the user accounts we know now, allowing storage of any activity on the connected service. The advantages this collection provides for these services is twofold: the first commercial personal content recommendation services were able to be implemented utilizing personal usage information. Services such as Amazon, for example, started linking relevant products based off activity history on their website, looking at purchase information and product views. In line with this is, secondly, advertisement personalization. Employing the same history, this could help to both improve the effectiveness through user specific adds, and therefore increase profits. Much of the information used for these advertisements relies on demographics: whereas search engines record queries, click-throughs, and location, social media sites stimulate storing even more personal information such as date of birth, addresses, relationship status, schools, life events, photographs and network of friends. Personal devices such as laptops, smartphones and tablets with integrated cameras, microphones and location data with constant access to the Internet makes connecting these personal pieces even easier. It provides the possibility for online services to build a very accurate description of what their users might be interested in, as they have a relevant part of personal profiles to their disposal. By combining all these profiles, it then becomes possible to discover trends and patterns in large collections of data, granting companies and services insights in their user base, from which for example marketing strategies can be induced.

As a result, the medium for information sharing became a medium of information analysis. Services with the largest datasets and most computation power had the biggest advantage as information became profitable. This, in turn, resulted in a skewed increase of corporate power, as Lanier (2013) argues. Companies such as Microsoft, Google, Apple, and Facebook all facilitated ease of communication and information sharing, and simultaneously built up vast amounts of detailed profiles of their users. Still, without these developments, and even the profits attached to them, the Internet would possibly not be as developed as it is now.

> *"If one would give me six lines written by the hand of the most honest man, I would find something in them to have him hanged."*
>
> – Bruce Schneier, citing Cardinal Richelieu

2.2 THE FREE INTERNET PARADOX

As we have seen, Net Neutrality states that the Internet is open, one is free to do as one wishes. For the longest time, this could be done anonymously; the standards and techniques used on the Internet were not focussed on profiling users yet. However, now, virtually any service-connected activity *can* be utilized in combination with your virtual identity. This realization through much of the debate surrounding Snowden makes it laden with the view and opinion Internet users have on privacy. The "Nothing to Hide" argument (Solove, 2011) is concurrently one much heard and many times refuted. However, the fact that this is a recurring opinion indicates that privacy is not a clear-cut concept, or right, that people believe they have, or everyone should have. The discrepancy in views and values between sides of this discussion might be better explained through reviewing previous research concerning privacy and primarily focussing on the shift that this has seen through time. The work of Smith, Dinev, and Xu (2011) provides an interdisciplinary review of research in this field, from which some important results will be discussed in this section.

2.2.1 *The Concept of Privacy*

The concept of privacy is a difficult, possibly impossible (Solove, 2006), one to give a overarching definition of. Different individuals might allocate a different value to it, based on history, their culture and governmental policies. It is therefore no surprise that research in the field of general privacy is broadly multidisciplinary. Still, as it is regarded to be nested in the moral values of society, the subject is mainly categorized as belonging to the study of ethics. Accordingly, many philosophers try to catch the concept in compact definitions, one of which is that of Solove (2004). In his book, he denotes them as being the following conceptions: (1) privacy as protection from Big Brother[4], (2) privacy as secrecy, (3) privacy as non-invasion, and (4) privacy as control over information use. Although Fuchs (2011) notes that many of these types of definitions are postulated, though not theoretically grounded, they do give a good sense of the shallow definition of privacy. It shows that privacy is in essence concerned with information,

[4] A popular synonym for intelligence agencies with a negative connotation, originating from '1984' by Orwell (1949).

and disclosure of this information, and more importantly a right to the ability of controlling these disclosures.

Similar theories can be found in the field of philosophy, including that of for example the *restricted access theory*, where privacy is considered a moral, transindividual structure, that is aimed at protecting humans. Another, the *limited control theory*, is often defined as a "*claim of individuals, groups or institutions to determine for themselves when, how, and what extent information about them is communicated to others*" (Westin and Blom-Cooper (1970) in Fuchs (2011)). This then assumes that privacy is dependent on human actions, where the individual controls what is disclosed and what is not. Giddens (2013) argues that the concept lies in the middle, in a combination of the two previous models named Restricted Access Limited Control (RALC). The underlying thought is that both society and individuals determine simultaneously what is disclosed, and individuals adopt their behaviour distinctively between their private life and the public, based on privacy and data protection. It clearly illustrates the divide that exists amongst more than just the field of philosophy. As we will see, the complexity of the concept clearly shows through in societal regulations of information privacy as well.

2.2.2 *Digital Privacy*

In Europe, digital privacy is regulated through the Data Protection Directive (DPD), offering protection recommendations of personal data. These seven principles (OECD, 1999), in short, grant full control over data that is collected on an identifiable person. A distillation of these principles indicates that this control is in the hands of that person self, which implies that when a third party is given enough transparency and purpose regarding the collection of this data through this person, that collection is deemed lawful. However, it could be stated that these DPD principles can only be fully effectuated if users are fully aware of the consequences their digital activity has. It is argued here, however, that the latter is not the case for many.

To illustrate, the actions of data collection and mining mentioned in the previous section might not have seemed daunting at first. After all, to a certain extent, users chose to share this information with the services they used. Although many of the digital footprints made online are not necessarily private, it can still be imagined that internet users would prefer not to have disclosed certain usage information to third parties which they, mostly uknowningly, still did and continue to do. This can be attributed to the fact that on many occasions, the acts of collecting and sharing this data are hidden behind pages of Terms of Service, and to some extent this camouflages them to uninformed users. However, even the majority of those with knowledge of these activities proved not to be bothered about this, with it having no noticable negative effect on them (Gross & Acquisti, 2005).

Of course, services using information to improve their own functionality could only be seen by one as beneficial. Furthermore, the increased sense of security when these methods of analysing information lead to tracking down 'cyber criminals' and potential terrorists are convincing and justifying to many people (Solove, 2011). However, through these surveillance actions, the line between a potential criminal and a regular individual has become increasingly blurred, especially with Internet Law being increasingly hard to effectuate. Technology is advancing faster than the legal systems of countries trying to impose Internet Law can effectuate, creating what is known as 'legal lag' (Rustad, 2011). Copyright laws, for example, have for many countries turned into a system of individual punishments of downloaders (Adermon & Liang, 2014), rather than that of the providers of this copyrighted content. Accordingly, websites such as The Pirate Bay[5] have up until now proven impossible to take down by court orders (Tweakers.net, 2013d). Moreover, real world concepts of privacy, such as secrecy of correspondence, do not seem to have an effective digital legal equivalent (Roba, 2009), as many online services, as we saw before, have explicit terms against them.

So through the use of online services, and disclosing information to these services, the aforementioned regulations implemented by the OECD (1999) that should protect digital privacy become ineffective. A user consciously utilizes and provides information to a service, and therefore gives a purpose to this data, relinquishing control over it. Whether this is a website such as Instagram[6] able to sell the pictures of their users to advertisement companies (Instagram, 2013), or the NSA granting themselves access to all theoretically collectible data; with the openness of the Internet and its rapid development also comes unavoidable exposure of information, and little to no means to protect information one would rather keep private. Maintaining anonymity has become increasingly hard, whilst many users willingly choose to share much of their personal information online.

Conclusively, this implies that the classic definition of the boundary between public and private domains has been blurred through the evolution of IT (Rosen, 2011). In this trade-off between privacy and usability on the corporate side (Chellappa & Sin, 2005; Hann, Hui, Lee, & Png, 2002), as well as privacy and security on the governmental side, lies a paradoxical relation. It shows that 'the principles of the Internet' are in effect as long as the users do not give these rights up themselves. Hence, the free and open Internet also underlies the complexity of upholding the concept of digital privacy. A logical conclusion drawn from many observations of the privacy paradox gave

[5] www.thepiratebay.se
[6] www.instagram.com

rise to the idea that privacy is currently a commodity: privacy is not an absolute individual and societal value; consumers, even when having high privacy concerns, still blindly provide personal information. It is therefore seen as a cost-benefit calculation for both the individual, and society (Bennett (1995) in Smith et al. (2011)).

2.2.3 *Privacy Concern*

The Free Internet, nor the digital privacy paradox are an axiom, or an inescapable by-product of the inherent aspects of our information society. Rather, it could be seen as a corollary of neglecting and misusing much of the legislations surrounding privacy, in order to achieve a set of corporate and governmental goals that are not in accordance with the aspect of human rights that these legislations hold. More importantly, as mentioned before, this is seemingly done with the internet users their passive consent. The paradox will be in effect as long as effective global legislation on access to and collection of data, and through this a strong framework for privacy, is implemented. The latter proposition lies at the core of the Snowden debate; with the disclosure of global surveillance, it can be hypothesized that to some degree, existing privacy concern has increased. Furthermore, it would be naive to assume that the passive consent to providing digital information is uniform.

Since the nineties, there has been conducted a fair amount of research in privacy concern, and the factors that induce or stimulate it, which logically sprung during the upswing of social media around 2005. Central to privacy concern in this field is the desire to keep personal information in control, away from non-related parties, and thus being able to communicate without interference (Buchanan & Paine, 2007). From around the noughties onwards, many polls and surveys indicated an increase in privacy concern, with percentages stable between 70% (Jupiter Research, 2002) and 72% (Consumers-Union, 2008) of American citizens being concerned that their behaviour was tracked by companies, whereas academic research in this field reports the same trend (Dutton, Genarro, & Millwood, 2005; Forrester Research, 2005; Harris Inc, 2004 as cited in Joinson & Reips, 2010).

Given this, it is peculiar that there is an abundance of research that proves that these attitudes do not translate into actual privacy-protectionist behaviour (Forrester Research, 2005; Jupiter Research, 2002; Metzger, 2006). For example, and as noted before from Gross and Acquisti (2005), Terms of Service (ToS) pose an at many times ignored free pass for online services to circumvent privacy legislations. Concern for ones privacy could be associated with informing oneself with the contents of the ToS, especially the privacy policy part. However, it has been shown that many ToSes are complex documents (Jensen & Potts, 2004), that are with few exceptions not read at all, and if read, frequently not understood (Berendt, Günther, & Spiekermann, 2005; Milne & Culnan, 2004). An additional example is that of users providing sensitive purchase-related information on a mock e-commerce website (Metzger, 2006; Spiekermann, Grossklags, & Berendt, 2001).

The research of the underlying factors of privacy concern could provide an explanation for this disjunction (Bellman, Johnson, Kobrin, & Lohse, 2004; Drennan, Sullivan, & Previte, 2006; Sheehan, 2002). From these, some interesting observations can be drawn: exposure to, and awareness of digital privacy risk raises suspicion, which subsequently results in active protective behaviour. This corresponds to results found by Buchanan and Paine (2007), who confirmed that students with a higher, technical education, were more cautious and used technical protection more frequently. From there, the hypothesis can be drawn that exposure to information on the topic of digital privacy, and the inherent problems we have discussed before, has an effect on awareness. This will be referred to as the *topic exposure* effect.

Additionally, people with an Information Technology (IT) background were found by Buchanan and Paine (2007) not to have increased privacy concern; however, they were deemed to be more aware of digital privacy risks. This can be explained by the fact that through both education and work, they have been more likely to come into contact with these risks. Accordingly, it could be argued that two groups can be distinguished: the *experienced* group, which are those with an IT background as discussed before, and the *non-experienced* group. The latter should be interpreted as follows: they are not assumed to be completely devoid of any technical knowledge. Rather, they characterized by the absence of a technical education, or special interest in the field, and are therefore deemed less likely to be knowledgeable about the many issues of digital privacy stated above. The chances of topic exposure for this non-experienced group are less mediated through interest and more through public coverage of the information.

2.2.4 *State of Dutch Privacy Discussion*

On the Dutch governmental of discussion surrounding Snowden and the NSA disclosures, an initial response by the Chamber of Representatives (Tweede Kamer der Staten-Generaal, 2013) articulated being committed to highly meticulous and adequate protection of personal data. Quoting their response: "*where national security and privacy protection meet, maximum transparency about procedures, powers, safeguards and oversight measures is a necessity*". Moreover, members of political parties such as D66, PvdA and VVD were making clear that (un)targeted dragnet data collection should not be something the government wants. In line with this statement is the no-confidence motion minister Plasterk received after hiding details surrounding the provision of 1.8 million meta data entries, containing phone calls, SMS and faxes, to the NSA (Tweakers.net, 2014b).

Meanwhile, however, the Dutch government has concrete plans to, similar to the NSA, intercept, collect and target search large quantities of data (Tweakers.net, 2013c). The intent to grant the Dutch intelligence agencies AIVD and MIVD more clout can be deduced from new legislations of wire tapping communication via cable, which is still prohibited by law, and the plans for Argo II, a social intelligence tool for which the technical specifications are classified. Despite objections by the digital civil rights organisation Bits of Freedom[7], the ministry has not proven to be keen in responding on the matter (Tweakers.net, 2013b). With no current response to the revision of data retention laws, and digital piracy having been illegalised by the European Court on top of this, the Dutch government cannot be clearly assessed on their positions regarding digital rights and the future of the Internet in The Netherlands. Given this, it might seem appropriate to conclude that, despite the fact that the Dutch government seemingly wants to invest in an intelligence agency harnessing similar functions to that of the NSA, public debate surrounding the topics this includes tends to be avoided.

Accordingly, one might draw the conclusion that the "post-Snowden era" has not (yet) affected actions taken by the Dutch government to mitigate the image of an emergent surveillance state. However, it is also a necessary conclusion that the complexity of political, as well as social implications behind the increasing collection of (private) information are well beyond the scope of this thesis. Still, given the fact that several (parts, or members of) political parties, as well as several organisations in The Netherlands have clearly voiced against these developments, along with Dutch newspapers having given these developments a sufficient amount of coverage as well, it might seem likely that the public opinion of both the experienced and the inexperienced group has developed over the last two years.

[7] www.bof.nl

> *"That's all it is: information. Even a simulated experience or a dream; simultaneous reality and fantasy. Any way you look at it, all the information that a person accumulates in a lifetime is just a drop in the bucket."*
>
> – Batou, *Ghost in the Shell*

2.3 RESEARCH QUESTION

In the previous sections the debate surrounding Edward Snowden was discussed, as well as both causes and factors of the inherent topic of digital privacy. It was noted, despite countries such as the United States having engaged in public debate on this topic, that in The Netherlands this seems to be kept to a minimum. In contrast, evidence was found regarding the fact that there has been enough potential for discussion, presumably away from the public, still having manifested. Given the discussed issues that underlie digital privacy, it is deemed likely that discussion on this topic has been relatively active prior to Snowden his actions. Even more importantly, it was hypothesised that discussion and concern surrounding digital privacy is directly influenced by both topic exposure, and topic experience. As such, it was argued that internet users with an IT background were more likely to have been aware of the possibility of privacy violation and the benefits intelligence agencies could have from harvesting the increasing amount of data before Snowden confirmed this. From there, it can be hypothesized that the media coverage surrounding this incident, over more than a year of releasing documents, has increased topic exposure and spread awareness to the general public, which might thus have triggered an increasing amount of opinions and discussions in the inexperienced group as well. It could then be argued that in line with the prior topic exposure the experienced group likely had, the inexperienced group was impacted more heavily by this news and accordingly, the amount of discussion can be expected to have increased significantly for this group. Hence, the thesis is that topic exposure through news surrounding the NSA increased overall privacy discussion, and influenced both groups differently as a result of their varying topic experience. To test these hypotheses, this research will aim to answer the following questions:

- To what extent has the Dutch discussion on digital privacy increased?
- To what degree can the difference in topic experience be observed prior to the leaks?
- How did these leaks affect both groups respectively?

Identifying and assessing this discussion requires a large dataset from a source that is able to capture developments of the discussion over time. From this dataset, snippets of discussion that are relevant to the topic set in the theory above should be able to be extracted and quantified. To employ this, computational methods from the field of Information Retrieval, as well as Machine Learning are required. Theory on automatic identification of topics is therefore an essential part of this thesis. First, however, it is required to get a sense of the Information Retrieval field, and more importantly document statistics, in order to understand this material. The next chapter will provide all information required to understand the method of the research posed, as well as certain design choices made in the dataset in.

"War... War never changes."

– Fallout

3 COMPUTATIONAL APPROACHES

Having stated the thesis, and defined the research questions, this chapter aims to provide the required background information on research methods from a combination of several fields that allow quantification of textual data. Section 3.1 will provide an overview of basic methods from the field of Information Retrieval through which word frequency and document similarity can be computed. More importantly, it will illustrate the complexity of retrieving topics from this data. Once these are defined, an effort to tackle the posed issues is discussed through an introduction of Topic Identification in Section 3.2. This will then lead up to discussing Topic Models, specifically Latent Dirichlet Allocation in Section 3.3.

3.1 INTRODUCTION TO DOCUMENT STATISTICS

So far, it has been established that part of the research aim is to quantify discussion on the topic of privacy. More importantly, it is preferred to be able to assess how the Dutch privacy discussion has developed over time. This implies that there is a need for interpreting large amounts of textual data containing these discussions, and a method of extracting information from this text. For the current research, this boils down to the ability to determine what the topic of a given document, or a set of documents is. Alas, few documents come with proper labels that state the exact topics it covers. Given this, there has to be found a method of using information from the text to extract these.

3.1.1 *Keyword Retrieval*

An initial solution could be to find a way to filter important keywords from the text, based on the frequency of their occurrence (Tokunaga & Makoto, 1994). Formally, given document d we would want to find the f occurrences of a unique word, or term t and store each of these in a list $\text{tf}(t, d)$, so that t with the largest frequency $\arg\max_t \text{tf}(t, d)$ could form a topic. Moreover, it would be preferred to filter out words that have no actual use in terms of being a topic, such as function

R	WORDS	TF
1	worden	7
2	teeven	6
3	gevolgd	4
4	burgers	4
5	mensen	3

R	WORDS	TF
1	teeven	32
2	bescherming	15
3	persoonsgegevens	14
4	staatssecretaris	12
5	privacy	12

Table 1: The rank (R), word, and tf for an article (in Dutch) on the privacy concerns that come with wifi-tracking (left), and for a set with six additional news articles on the topic of Fred Teeven (right).

words, adverbs, articles, etc. Keeping the sanitation simple for the sake of this example, an article that is labelled to be on the topic of privacy (Tweakers.net, 2014a) is taken, and words shorter than five characters are filtered. This results in a *term frequency* (tf) list from which the top five words can be found in Table 1.

When observing these results, however, there is no evidence of its topic amongst the five most frequent keywords. In fact, the word `privacy` itself was not in the article at all. This is referred to as a *latent* topic; it is not directly derivable from the text alone. Alternatively, it can be deduced from the fact that (Fred) `teeven`, the Dutch state secretary for Security and Justice, can be found in more news articles that directly link to privacy-related topics. Adding six related documents (Tweakers.net, 2010a, 2010b, 2012a, 2012b, 2013a, 2014c) and summing their frequencies yields the results shown in Table 1, where both `bescherming persoonsgegevens` (personal data protection) and `privacy` evidently score higher. This clearly illustrates a problem of content richness; one news article is short, a couple of paragraphs at most, and therefore does not provide enough information on its own to accurately determine a topic. Combining the information from documents that related, however, seems to create some collective representation where a topic can be deduced from. Still, the related documents were picked by hand in the example; therefore, there still needs to be found a method to determine which documents are related. Luckily, this need for information has been extensively discussed in the field of Information Retrieval, which has rapidly enhanced the task of statistical analysis on documents with a wide variety of computational models. With the background provided by this section, some of these can be discussed.

TERM	D1	D2	D3	D4	D5	D6	D7
`teeven`	0.000	0.000	0.000	0.000	0.000	0.000	0.000
`bescherming`	0.146	0.000	0.146	0.146	0.216	0.000	0.286
`persoonsgegevens`	0.359	0.000	0.243	0.243	0.000	0.000	0.475
`staatssecretaris`	0.067	0.107	0.099	0.067	0.067	0.000	0.067
`privacy`	0.359	0.000	0.243	0.000	0.359	0.389	0.000

Table 2: A tf · idf weighted term · document matrix for the set of articles on the topic of Fred Teeven.

3.1.2 tf · idf

Defining the seven articles used in Table 1 as a collection of documents $D = \{d_1, d_2, \ldots, d_N\}$, where accordingly $N = 7$, it can be evaluated how important some term t in each of these documents is given the collection. The rationale behind this is that terms with a high tf in the entire collection are deemed uninformative. To illustrate, given D, `teeven` will not be helpful in discriminating between the individual documents in this set. These high tf values can be reduced through the amount of documents in the collection that contain the associated t; the *document frequency*, denoted df. Using a logarithmic function, rare terms can then be scored higher than common ones through the *inverse document frequency*, denoted idf. These measures can be formalized as:

$$\mathrm{df}_t = |\{d \in D : t \in d\}| \tag{1}$$

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \tag{2}$$

Combining the definitions of tf and idf results in a method of weighting each term, in each document, called the tf · idf weight. Formally:

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t} \tag{3}$$

The application of this weight to the example set can be found in Table 1. It can be observed that the previously high scoring `teeven` has now become unimportant over the entire set, as it is not characteristic for any $d \in D$. Moreover, topics such as `privacy` prove to be more characteristic for some documents (D1, D3, D5, and D6) than others. With this measure it is also possible to quantify a similarity between documents that might then contain the same topic.

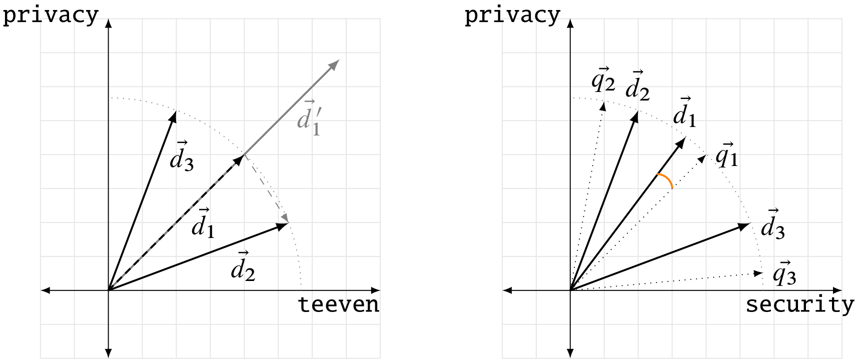


Figure 1: Two-dimensional Vector Spaces, showing document vectors $\vec{d}_1$, $\vec{d}_2$, $\vec{d}_3$, and $\vec{d}_{1'}$ to the left. To the right, document vectors $\vec{d}_1$, $\vec{d}_2$, $\vec{d}_3$, as well as topic vectors $\vec{q}_1$, $\vec{q}_2$, $\vec{q}_3$ are displayed.

3.1.3 *Vector Space Model*

The set of tf · idf weighted documents can be seen as vectors $\vec{V}_d = (w_{d,1}, w_{d,2} \dots w_{d,t})$ where each $w_{d,t}$ is a weight for term t in document d. Accordingly, the contents of these vectors look similar to those in the columns of Table 2. This representation of documents as vectors is known as a Vector Space Model (VSM) (Salton, Wong, & Yang, 1975). Note that each document vector has a *bag-of-words* representation; it does not concern itself with the position of any word in relation to another. This implies that `Hank and Marie` is exactly the same sentence as `Marie and Hank` in the eyes of the model, as the words are thrown in a bag and picked without looking, so to say.

Each of these $\vec{V}_d$ denotes a direction of t per d and magnitude, or length, $||\vec{V}_d||$. In principle, each number in a list of vectors represents a direction in a certain dimension, with the term axes being the *basis vectors*, defining the vector space. The membership of the different vectors to a certain term can be quite intuitively perceived by looking at the two-dimensional VSM in Figure 1. While $\vec{d}_3$ is closer to the `privacy` axis, indicating more similarity to that topic, $\vec{d}_2$ is more closely related to `teeven`. In reality, however, a VSM is very likely to have more $|V|$ dimensions, through a manifold of documents and terms, than humans can visualize. This two-dimensional example should give some notion of the space in a simplified fashion, however.

3.1.4 *Cosine Similarity*

So, with these vectors an attempt can be made to quantify the similarity between different documents (Manning et al., 2008), for which the VSM is used. This might be done by comparing the magnitudes of their vector difference. However, the problem is that documents are apt to differ in size and therefore result in a longer vector as indi-

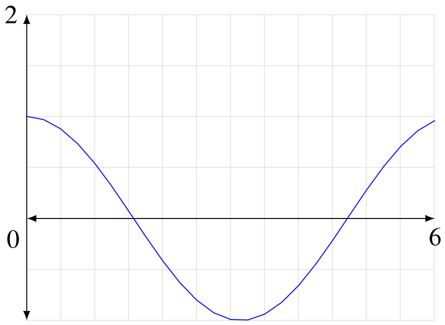


Figure 2: The Cosine function.

cated by vector $\vec{d}_1'$ in Figure 1. Consequently, this will result in a large distance between different vectors of different length. To compensate for this, vectors such as $\vec{V}(d_1)$ and $\vec{V}(d_2)$, can be length normalized, reducing $\vec{d}_1'$ to $\vec{d}_1$ in Figure 1. For the similarity between documents, this normalization is commonly represented in a *cosine similarity* measure (sim). The cosine function decreases monotonically between 0 and 180 degrees, as can be seen in Figure 2, and is therefore the inverse score of an angle. Formally:

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} = \vec{v}(d_1) \cdot \vec{v}(d_2) \tag{4}$$

Two products are calculated here: Firstly, the *dot product* of the vectors $\vec{V}(d_1) \cdot \vec{V}(d_2)$ defined as $\sum_{i=1}^{M} x_i y_i$ where M is the number of components in $\vec{V}_1(d) \ldots \vec{V}_M(d)$. Secondly, the *Euclidean length* of d, or $\sqrt{\sum_{i=1}^{M} \vec{V}_i^2(d)}$, which length-normalizes the vectors. With this cosine similarity, it is possible to rank documents based on how relevant they are to a certain term t in a vector space.

3.1.5 *Document-Topic Similarity in Practice*

For a practical example, consider $t_1 =$ `privacy` and $t_2 =$ `security` as the axes of a VSM, represented in the right space of Figure 1. To determine the topic of the documents in this space, three topic vectors are added: $\vec{q}_1 =$ `RSA encryption algorithm`, $\vec{q}_2 =$ `RFID chip`, and $\vec{q}_3 =$ `TrustyCon`, where $\vec{q}_n = (w_{t,1}, w_{t,2} \ldots\ w_{n,t})$. Now, rather than calculating the cosine similarity between given d_1 and d_2, the similarity between topics and documents is calculated: $\text{sim}(d_n, q_n)$. With this similarity (illustrated by the arch between $\vec{d}_1$ and $\vec{q}_1$ in Figure 1) it can be determined that $\vec{q}_1$ is most similar to $\vec{d}_1$, from which it could then be deduced that d_1 has `RSA encryption algorithm` as a likely topic candidate. So, given some predefined range of topics it is possible to locate the most likely candidates for these topics within the VSM.

The crux here, however, lies in the fact that measures such as tf · idf are keyword based and do not capture the *latent* topics discussed before to the extent that would be necessary to accurately capture a topic. Consider the following example:

$d_1 =$ `Tweakers.net is a nice website.`

$d_2 =$ `I visit Tweakers.net on a daily basis.`

$d_3 =$ `Tweakers.net is a tech website.`

$q_1 =$ `Tweakers.net recommendation`

Multiple problems surface here: using a VSM[1] to determine the cosine similarity between q_1 and (d_1, d_2, d_3) will result in the documents d_1 and d_3 being closest in the vector space, as there is only one non-overlapping word between them. Intuitively, however, one could argue that, in light of the stated q, d_2 is more similar to d_1 than d_3 is; they both show a positive attitude towards the website `Tweakers.net`. Moreover, the topic q_1 only matches with the first word onto all documents, and therefore results in an equal similarity score for all of the documents. Obviously, the model has no knowledge of the meaning of the word `nice`, or regarding the fact that visiting a site on a daily basis would imply that one would likely recommend it. As such, it cannot infer these documents are a `recommendation`, as it cannot make this kind of connection without further context.

kill	…	…
…	double	…
…	…	base
…	gunned	…
ammo	…	…

handheld	…	…
…	screen	…
3d	…	…
…	stylus	…
…	…	battery

…	…	guild
…	elf	…
spell	…	…
…	talent	…
tree	…	…

town	…	…
…	…	sword
princess	…	…
…	song	…
…	…	horse

Table 3: Bag of words representation of a set of documents.

[1] The models that will be discussed in the following sections all build (if not specified otherwise) on the tf · idf measure in this VSM; therefore, this should provide enough background knowledge to understand their theoretical foundations.

This keyword based retrieval also falls short when searching for documents with a similar theme, or topic. Despite the fact that in a large collection some documents might concern a certain topic, the absence of mentioning these topics explicitly might drastically change its position in the VSM and will therefore make them difficult to retrieve under one topic. Considering Figure 3 to be four documents in a collection, and naively assuming that these are the only words available for the analysis of these documents, one could choose to guess based only on the individual content of the documents what their topics are. This process could lead to the conclusion that the topics are respectively `army`, `gadget`, `fantasy`, and `middle ages`, which might seem appropriate topics at first glance; however, considering all these documents were in reality on the topic of `video games`, the interpretation of the document radically changes. Therefore, if it would be possible to capture these topics through the collective topical value the words they are associated with possess, one would be able to retrieve documents by actual topic rather than keyword. This is the exact information need that the field of Topic Identification deals with.

> *The leadings fine artist, the painter Hans Drucker, raised his hand and said, "young man" addressing Georg Nees, "all said very well, what you told us, but you know what, could you make your machine draw the way I draw?" and Nees pondered for a moment and said, "you know what, if you tell me how you draw I can make my machine do it."*
>
> – Frieder Nake, *Hello World! Processing*

3.2 INTRODUCTION TO TOPIC IDENTIFICATION

When large corpora and repositories of textual information became publicly available and integrated into information retrieval systems, more intelligent approaches to deal with the 'vagueness' of user requests were needed. Especially when issuing a search, or query to find something in a document, word choice might differ per person. Moreover, the mere content and word choice of a document might not accurately cover the actual query sufficiently. Even more problematic are *polysemy* and *synonymy*, which cause major deficiencies in classic retrieval approaches such as tf·idf. This is exactly what the introduction of Latent Semantic Analysis (LSA) (Deerwester, Dumais, & Landauer, 1990; Landauer & Dumais, 1997) attempted to solve. Through a method for statistically representing the contextual-usage meaning of words, the semantic space was introduced, whereupon many modern Topic Models are built. Given this, understanding LSA is an essential first step into understanding the field of Topic Identification, for which in depth overviews of the field by Steyvers and Griffiths (2007) and more recently Burtion (2013) provide a solid background, from which this introduction will draw. Accordingly, Section 3.2.1 will explicate LSA its required theoretical background. After, by introducing Topic Modelling in Section 3.2.2, a global overview of the field its history is provided, where the focus primarily lies on Probabilistic Topic Models as will be discussed in Section 3.2.3. This will then offer a solid background in order to grasp the notions behind Latent Dirichlet Allocation, which will be discussed in Section 3.3.

3.2.1 *Latent Semantic Analysis*

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) was introduced in 1990 as a system that, similar to tf · idf, does not use any extra knowledge aside from that of the raw text itself; however, unlike tf · idf, relies on meaningful passages to extract information from, rather than the full document. The strength of LSA lies in the fact that these passages are isolated through a process of matrix factorization, by applying (Reduced) Single Value Decomposition (SVD) (Baker, 2005) to a matrix of words and their relevance score in pas-

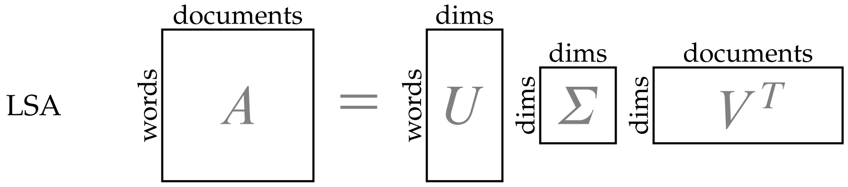


Figure 4: The matrix factorization of the LSA model (adapted from Steyvers and Griffiths (2007).

sages (recall Table 2). SVD decomposes[2] this matrix **A** into the product of three other matrices: an orthogonal matrix **U**, a diagonal matrix $\boldsymbol{\Sigma}$, and the transposition **V**. Formally:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \tag{5}$$

Breaking down the math into a more graphical model, as that of Figure 4, clearly shows how an initial term · document matrix is broken down into these three separate matrices, decomposing original matrix **A**. The essence of the process is that it identifies new dimensions in $\boldsymbol{\Sigma}$, along which instances show the most amount of variance, and are therefore least correlated in this matrix. Even more importantly, the orthogonal nature of these decomposed matrices allows dimension reduction through only retaining a top k highest singular values, after which the individual matrices can be reconstructed into a low-rank approximation $\mathbf{A}_k$ to the initial term · document matrix.

The result is a k-dimensional representation of the VSM discussed before, also referred to as a semantic space, where semantically related term · document pairs now have higher similarity. The rationale behind the model is, subsequently, that co-occurring terms within a document, will have a similar representation in this space, even though they might not share common terms, and can therefore be used to identify words that refer to the same topic. The latter makes this model lie at the heart of topic modelling and has formed an important starting point in the Topic Identification field.

3.2.2 *Topic Modelling*

The task of Topic Identification is that of discovering underlying topics in a set of documents by means of a topic model. Topic modelling, in turn, is a term for a set of computational techniques to find patterns

[2] To understand the terms and calculations used in this section, some relevant linear algebra theory is required, for which a brief explanation can be found in Appendix A. Appendix section A.4 discusses the steps that need to be performed in order to apply a Singular Value Decomposition to a given matrix **A**.

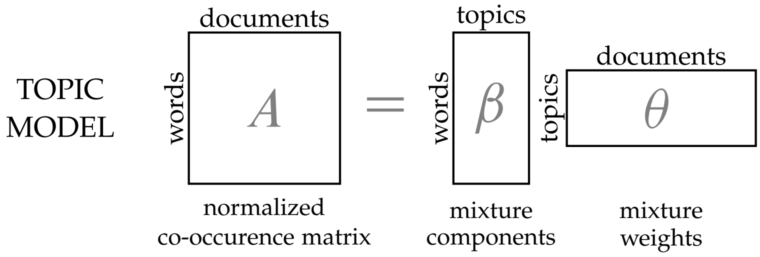


Figure 6: The matrix factorization of the topic model (adapted from Steyvers and Griffiths (2007).

of co-occurrence in this set of documents. Since the development of LSA in 1990 (Deerwester et al., 1990), the field of topic modelling has seen much improvement and a variety of approaches in different fields of research, with models such as pLSA (Hofmann, 1999), (S)LDA (D. Blei & McAuliffe, 2007; D. Blei, Ng, & Jordan, 2003), and HDP-DBM (Salakhutdinov, Tenenbaum, & Torralba, 2011).

As stated before, there is a conceptual similarity between LSA and topic models, as both reduce their input matrix to a low-dimensional representation. Looking at Figure 6 clearly illustrates these similarities; with a word · document matrix **A** being split into a topic matrix $\boldsymbol{\beta}$ and document matrix $\boldsymbol{\theta}$. However, this also simultaneously displays the differences between the two approaches: in topic models, the word · document vectors are distributions of membership probabilities, that accordingly sum up to one. The models are based on the idea that a document contains a mixture topics, in which a word has some probability of occurring. For example, when writing a document on the subject[3] of `android`, the word `HTC` from the topic `mobile phones` is more likely to be used than `iPhone`, and `open-source` from the `apps` topic more than `jailbreak`. In addition, there might be a higher chance of the first topic occurring in this than the latter, as will be discussed in Section 3.3.

Thus, in essence, a topic model is a generative model, as it tries to induce a model by which a document could, theoretically[4], have been written. Steyvers and Griffiths (2007) illustrate this using Table 4; with equal probabilities given to the first two topics, a document can be constructed regarding a patient its colour perception being affected by drugs, whereas the same for the last two would yield a document about memory loss treatment. These probability distributions under-

[3] Note that there is a difference between the real-world concept of a subject, and the language model definition of topic; a distribution over words.

[4] One of the more important notions within the field is that it does not draw from the natural language structure humans might infer context from, by means of word order, semantics and such. Rather, similar to a lot of other approaches in Natural Language Processing, it chooses to ignore structure in a text and to treat words as individual cases; this *bag-of-words approach* was also shown in Figure 3. This, most notably, implies that a text generated through a topic model is not a aesthetically pleasing piece of human writing.

Topic 247		Topic 5		Topic 43		Topic 56	
WORD	PROB.	WORD	PROB.	WORD	PROB.	WORD	PROB.
drugs	.069	red	.202	mind	.081	doctor	.074
drug	.060	blue	.099	thought	.066	dr.	.063
medicine	.027	green	.096	remember	.064	patient	.061
effects	.026	yellow	.073	memory	.037	hospital	.049
body	.023	white	.048	thinking	.030	care	.046

Table 4: Four topics from the TASA corpus.

lying the generative process are, thus, what defines a topic in this sense; a distribution over words. However, this generative process can be utilized the other way around as well; to infer a set of topics that are underlying a collection of documents. For this thesis, the interest lies primarily in the latter application, which is Topic Identification utilizing Probabilistic Topic Models.

3.2.3 *Topic Identification*

The field of Topic Identification focusses on employing these topic models, based on the probability distribution, to identify one or multiple topics in full texts, or snippets thereof. There are numerous approaches of integrating topic models into algorithms for classification (Phan, Nguyen, & Horiguchi, 2008; Quercia, Askham, & Crowcroft, 2012; Ramage & Hall, 2009; Srivastava & Sahami, 2009); however, only the intuition behind this classification will be discussed in this section.

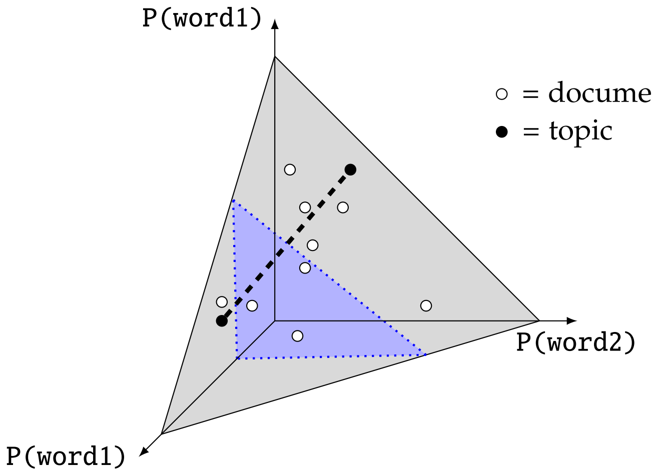


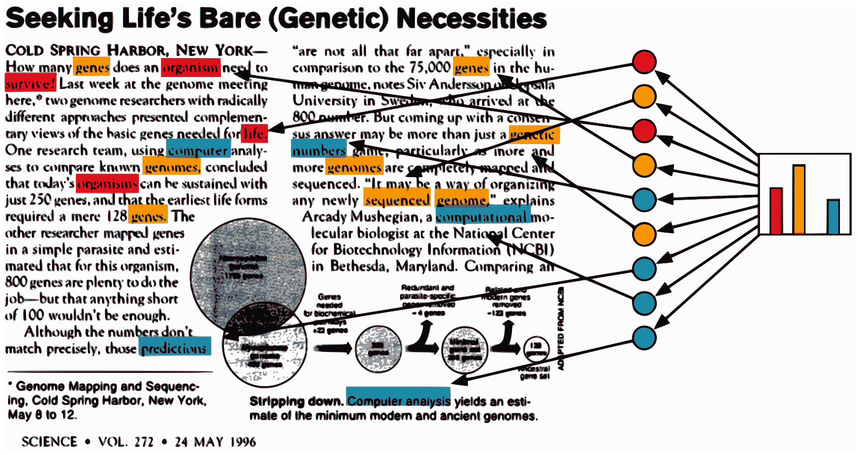
Figure 8: A geometric intepretation of the topic model, adapted from Steyvers and Griffiths (2007).

Given that a topic t is a distribution $P(w|t)$ over words w, $P(t)$ forms a distribution over topics in a particular document. Taking a vocabulary with distinct word types W, these probabilities can be translated into a W dimensional space (recall the VSM). The axes, as in Figure 8, are then a probability of observing a certain word. Given this, each document can thus be represented as a white point on the grey simplex, by the weight that each word has in the document. Therefore, topics can also be placed in this space, represented as a black point. With this space, it is possible to create a linear classification problem that is employed in the field of Data Mining. From there, a decision boundary that divides the documents into being, for example, either a member of topic A, or topic B can be determined. In Figure 8 this is illustrated in a simplified manner by the dotted, smaller simplex. In reality, there presumably are numerous more topics and words, and therefore more dimensions than can be visualized. Creating these decision boundaries can be done with a range of classification methods such as for example Naive Bayes, k-NN, SVM (Witten & Frank, 2005) or MaxEnt (Berger, Pietra, & Pietra, 1996). Other approaches to classification will be further explained in the method chapter; however, first a more detailed look will be taken into a more advanced Topic Model; Latent Dirichlet Allocation.

> *"I just set the number of topics here to fifty, I didn't strike the right balance carefully. It's like I cook, the same way I cook. You know, I just put some chilly pepper in and it tastes like chilly pepper."*
>
> – David Blei

3.3 LATENT DIRICHLET ALLOCATION

The intuition behind Latent Dirichlet Allocation (LDA) (D. Blei et al., 2003) is that a document d exhibits multiple topics t. It is assumed that a collection D of d has a fixed vocabulary, and any t is a distribution over this vocabulary, with any word $d_w \in d$ having some probability in t as t_w, and having higher probabilities for the t they are most associated with. So in essence, a topic t is defined as a distribution over words, such as those in Table 4. This idea is then cast into a generative probabilistic model, which can explain why parts of the data are similar. LDA chooses a probability over topics in a document, as illustrated in Figure 10. From this space, a distribution over words can be chosen, and then a word from that distribution, which is repeated for each word in a document, for each document in a collection with a different distribution for each document. Recall that, similar to any topic model, this assumes that the order of the words does not matter, which implies that employing the generative model is not going to produce documents that make much sense; however, the goal of LDA is to find thematically coherent groups of topics, not to generate coherent text.



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive! Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Genes needed for biochemical pathways +22 genes
Redundant and parasite-specific genes removed −4 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure 10: A visual representation of the Latent Dirichlet Allocation topic modelling procedure by D. Blei (2010).

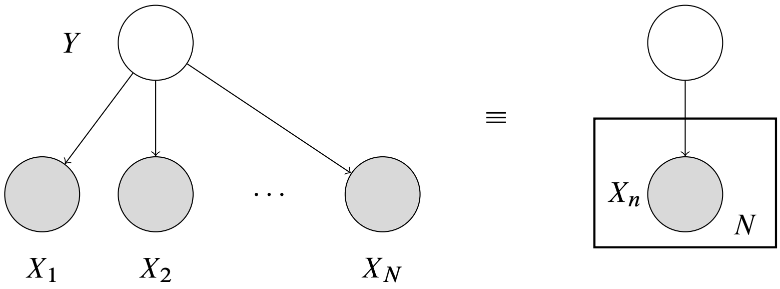


Figure 11: Basic graphic model (D. M. Blei, 2009).

3.3.1 *LDA Formalized*

This process can be illustrated in a more graphic and formal way (Figure 11). Given the grey, observed variables $x_1, x_2 \dots x_n$ a latent concept Y can be inferred. This is notated as, and therefore $\equiv$, a boxed N number of replications for all observed features, which is called a plate notation. It therefore defines a pattern of conditional dependence between all variables in X_n, in this example corresponding to:

$$p(y, x_1 : x_N) = p(y) \prod_{n=1}^{N} p(x_n|y) \tag{6}$$

Looking at Figure 12 shows that choosing the distribution over the topic collection plate K is a repetitive process, where β is a distribution over words (a topic), with a k amount, and β_k resides in a simplex; the space of all possible distributions.

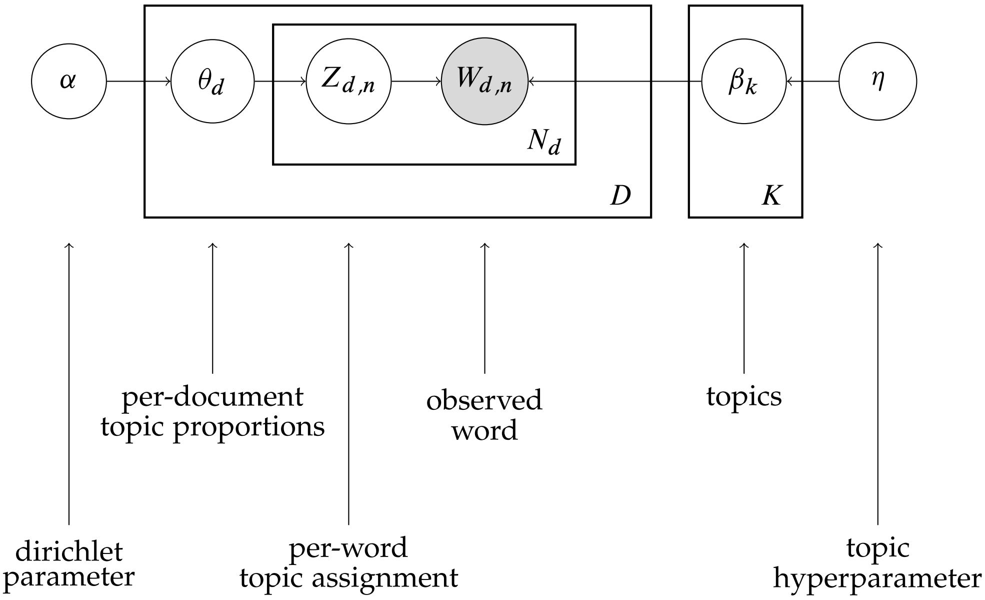


Figure 12: The plate notation for LDA (D. M. Blei, 2009).

Then for each document in the document collection plate D, θ_d is the coloured histogram as seen in Figure 10, the distribution over individual topics is chosen randomly from an overarching distribution of distributions, also called a Dirichlet. Then, for each word N_d in D, a topic indicator $Z_{d,n}$ (the topic number) is chosen from this distribution with parameter θ_d, which is illustrated in Figure 10 as the coloured circles. Note that $W_{d,n}$ relies on both $Z_{d,n}$ as well as all β_k, as it is the n-th word in the d-th document, and it is the only observed variable here. Therefore, given a topic matrix such as Table 4, the $Z_{d,n}$-th column with the $W_{d,n}$-th word is selected from β_k,and the probability is taken from there. So, say that topic 43 will be chosen, the 43rd topic is looked up, after which a word from that topic is drawn. This would formally be written as:

$$\left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d)p(W_{d,n}|Z_{d,n},\beta_{d,n})\right)\right) \tag{7}$$

3.3.2 *The Dirichlet Distribution*

The Dirichlet represents a distribution over some number of elements in the simplex $K-1$ (D. Blei et al., 2003); positive vectors that sum to one, formally:

$$p(\theta|\vec{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1} \tag{8}$$

The parameter values for α are drawn from a vector of length K, with $\Gamma(x)$ being the Gamma function, assuming that $\theta \sim \text{Dir}(\alpha)$. Consider having a uniform distribution with three parameters for α being $\theta \sim \text{Dir}(1,1,1)$. Drawing from this θ will result in distributions from three (a, b, and c for the sake of this example) different elements. Depending on the weights in the distribution over these three different elements that make up α, it will result in different points. Placing all mass in a, for example, as in the middle bars of Figure 13, will result this point to be very close to a in the simplex.

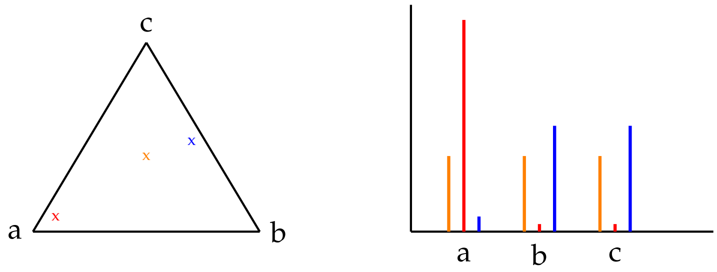


Figure 13: The simplex position of three different distributions over elements a, b, c.

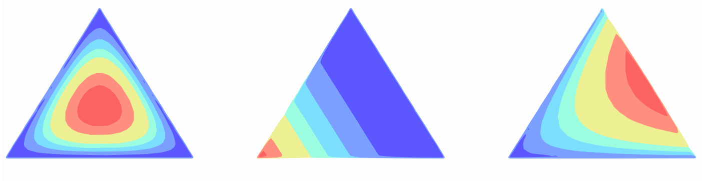


Figure 14: The Dirichlet Distribution.

The left bar has its mass divided over all elements and is therefore in the centre. Now assuming that we do not use the uniform distribution $\theta \sim \text{Dir}(1,1,1)$ with $\alpha_1 = \alpha_2 = \alpha_3 = 1$, but all equalling 5 for example, the probabilities over these distributions will cause a bump; the inner level of the gradient shown in Figure 14. The location of the bump is determined by the expectation $E[\theta_i \,|\, \alpha] = \frac{\alpha_i}{\sum \alpha_i}$ which would be $\frac{1}{3}$ given $\alpha = 5$ $(\frac{5}{15})$. The larger $\sum \alpha_i$, the more peaked and therefore less spread out the bump will be at its E. This is mirrored by the following: say N topic assignments are observed, what is the conditional distribution of the proportions given those topic assignments? With $n(Z_{1:N})$ (with $1:N$ denoting 1 through N (D. Blei et al., 2003)) being the counts of each topic, and topic 11 occurring six times, then:

$$p(\theta \,|\, Z_N) \sim \text{Dir}(\alpha + n(Z_{1:N})) \tag{9}$$

The amount of times topic 11 occurred is then added to α which is then the posterior Dirichlet. So as more data is observed, the model becomes more confident about the distribution and therefore the Dirichlet will peak more.

3.3.3 *Topic Modelling with LDA*

Conclusively, LDA produces topical words, placing high probability on words that co-occur because co-occurrence, and the dependency of the words that co-occur, are linked. Thus, to have a certain topic implies that the words within this topic are a logical dependent of each other in forming a text with this topic (D. M. Blei, 2009), while the word probability is maximized by putting them in different categories (Figure 7). Blei illustrates that when using a mixture model,

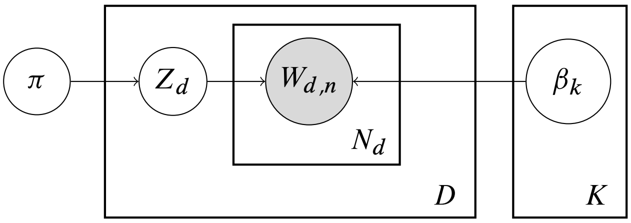


Figure 15: Mixture Topic Model where topic assignment Z_d only relies on the document where it is from (D. M. Blei, 2009).

it is assumed that one document has one topic, so each word in that document belongs to some topic X (Figure 15). Such a model will exactly capture sets of recurring, co-occurring words. If multiple words occur in different topics with a high occurrence, it will be very problematic for the model. In contrast, when using LDA, a document will be penalized for having too many topics through the Dirichlet, encouraging sparsity. This implies that the model prefers topics that only have some elements in their distribution with a high probability, as well as LDA being transformed into a mixture model when the value of α is very low. This allows LDA to accurately model a document X containing topic A, as well as topic B independently, without polluting information on topic A and B respectively or creating a document that exactly fits document as an AB topic. For example, given a topic on `privacy`, as well as `android`, these need to be identified as separate topics, so it will not incorporate words such as `encryption` co-occurring in our `android` topic, which would happen in a simple mixture model. Through this, LDA hopes to strengthen sets of terms that more tightly co-occur. LDA builds on the kind of models shown before; namely LSA (Deerwester et al., 1990) and PLSA (Hofmann, 1999), and is therefore considered a mixed membership model due to the fact that Z is not associated with a single document, but a distribution over multiple through θ.

3.3.4 *Labelled LDA*

What makes LDA important for this research is the fact that its supervised implementation, Labelled LDA (L-LDA) (Ramage & Hall, 2009), is well documented and fairly easy to implement. In a way, it constrains LDA through a one-to-one correspondence between the latent topics and human-provided topics. In the model in Figure 16 this corresponds to the restriction of θ_d to only be defined over the topics that correspond to Λ_d, with Φ_k being the label prior for topic k. This way, a multi-labelled set of documents can be used for learning the topic distributions, which is generally done through optimizing the model's performance with Gibbs Sampling (Griffiths & Steyvers,

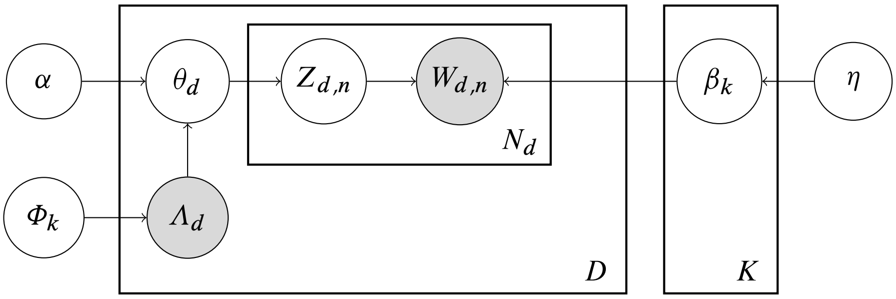


Figure 16: The plate notation for Labelled LDA (Ramage & Hall, 2009).

2004), and inferring topics for documents in a new set. In this way, L-LDA allows classification and therefore Topic Identification, as was discussed in Section 3.2.3. Now that it is determined which computational approach will be utilized for topic classification, the required dataset can be determined through assessing both the research questions from Chapter 2, and the supervised model stated here.

"So no, I'm not required to be able to lift objects weighing up to fifty pounds. I traded that for the opportunity to trim Satan's pubic hair while he dines out of my open skull so a few bits of the internet will continue to work for a few more days."

– Peter Welch, *And Then I Thought I Was a Fish*

4 METHOD

Within the framework of the theory discussed before, it has been established that the goal of the thesis is to assess and identify the Dutch discussion on digital privacy. To employ this, a computational model for topic modelling was chosen. Given this, it can be determined what the dataset should contain. Before constructing the actual dataset, however, it is important to carefully select the source(s) from which it will be formed. This chapter, therefore, will first explicate how the choice of source type is established. Secondly, requirements concerning the sources of choice will be explained, along with the limitations they bring forth. After, the sources will be briefly introduced in line with these requirements.

4.1 DATASET

While discussing the computational approach to classify topics, examples of textual data were looked at. What is described here is how a set of such documents is established for the purpose of research on discussions. Therefore, some of the terms that have been used before will again be formalized to create an understanding how those will relate to the dataset. Moreover, the sources to obtain these documents have a key role in successfully performing this task. As such, any decisions regarding the selection of these will be discussed.

4.1.1 *Data Requirements*

To clearly structure the data requirements, an overview of the several factors that impose these requirements is given. These are noted in a top-down representation, where the *research*, followed by the *dataset* and the *documents*, as well as the *source* make up these factors. Any terms that will be used throughout the research in the way they are formalized here will be emphasized.

THE RESEARCH requires *documents* that can provide a clear *timeline* as well as enough *content* to deduce their topics from. The term document, here, refers to a collection strings of text that are related by the topic they describe. In this sense, a book or a page on Wikipedia that focusses on a certain topic is an example of what is regarded as a valid instance of the term document. These strings of text supply information regarding the topic, referred to as *content*. The more information is provided within a document, the more links can be drawn to various related topics, the more content rich a document becomes. The requirement of a timeline implies that the document itself, and perhaps more preferably its parts, are labelled with timestamps, providing information on what date and time these were added.

THE DATASET requires documents that contain streams of *discussion* that are facilitated within an active *community*. Discussion, then, entails written conversations that are individually intended as a form of reaction on the document. Imagine, for example, an online blog post as document, with the reactions forming the discussion stream. These could for example contain an opinion on the topic or an addition on the subject in terms of information. Note that these reactions might or might not be directly in line with the topic of the actual document, as long as they are clearly linked to the instance of the document, they are deemed valid. The persons producing these streams of conversation, in turn, can be seen as a group, which is in turn considered a community here. The community might not be the author of the document's content; as it could well be written by persons not participating in the stream of discussion. However, the community produces the most relevant addition to that document for this research, namely the user generated content they provide on the document instance. Consequently, the more active the community, the larger the amount of discussion and therefore the higher the chances that discussion itself has more content and varying topics within one document.

THE DOCUMENTS must be easily, and thus openly, accessible and allow for constructing a dataset of a decent size. Roughly said, the research requires quick access to a large amount of data. Fortunately, the Internet allows for searching and collecting suitable documents and, more importantly, do this in an automated fashion. Obviously, not just any random website can be picked to collect the documents from. It was already established that the content of these documents must be susceptible to extensive discussion. Online communities, groups sharing and discussing information on the Internet, therefore logically pose a fitting source to meet the established requirements. However, most of these communities handle topics in an ad-hoc fashion. Taking fora (online discussion boards) for example, a topic only gets initiated when one of the users decides to create a thread discussing it. Blog sites might pose the same issue; despite the fact that the initial post might be more content rich than that of a forum, to some extent, the interaction and comments are assumed to be more limited.

THE SOURCE taking the above into regard, would seem to fit the requirements imposed by the theoretical framework to a greater extent if it represents a more stable and regulated source of information. Accordingly, online communities found on news websites fit these requirements, as they incorporate topics susceptible to discussion on a daily basis, as well as provide enough interaction (in terms of community size and contribution activity). This can be substantiated in fourfold: firstly, current affairs that are discussed in news articles are argued to likely form opinions and would therefore initiate more discussions in large communities than in the sources mentioned before. Secondly, the larger, therefore attracting more visitors, the website, the more (qualitative) articles are likely to be posted on a daily basis. This is even more true when the articles are based on an official news source, which one might expect to be referred to more often with popular news sources. Thirdly, as the date and topicality of news is important, the articles can easily be placed on this timeline posed as a requirement before. Accordingly, certain topics can be divided amongst multiple iterations of an event. As such, they impose a certain continuity with developments within these events on the discussion per article. Finally, these websites often provide a structured framework wherein users can submit their discussion, which is generally referred to as the 'comment section'. The users of the community can utilize the section by, at most times, registering with a unique username through which they become identifiable. Each reaction, or comment, is stamped with a date and time. Moreover, often a kind of hierarchical structure is applied to comments that are a reply to a previous comment, from which individual discussions might be distilled. This to some extent simplifies retrieving and linking of this data. Considering all these advantages to the characteristics of online news websites, it is feasible to deem these as an ideal source for building the required dataset. This concludes the selection of source type; however, for the different possible news sites being selected, requirements must also be established.

4.1.2 *Source Requirements & Limitations*

Now that the requirements for the data, thus the source type, are established, the sources can be zoomed in on and subjected to a list of more specific requirements in terms of community, language and structure.

4.1.2.1 *Community*

Along with the generally large amount of content that the comment section of the candidate website(s) should provide, the characteristics of the communities also serve an important role in selecting news websites appropriate for the posed research. Limiting to these websites namely yields some complications that must be taken into regard while choosing the communities used for processing the content used to construct the dataset. Most of the wide variety of news websites do not allow visitors to discuss on their articles or do not have a community active enough to even provide more than ten comments on each article. Moreover, as discussed before, numerous news websites have a different content focus, ranging from general news, to specializing in sports, technology or for example science. Therefore, comparison between communities is generally difficult. Logically, these different news sites also attract different types of users. It can be imagined that a general news website will have a community that more resembles a general population intersection than a specialized website.

A side-effect of this topic differentiation might therefore be that users with a certain interest a generally more engaged in discussion when the topic pertains these interests. To illustrate, a website that predominantly focusses on vegetarians is more likely to have a vast amount of discussion on animal rights, whereas a website with no specific overarching subjects, though likely to have more visitors, would even total have less discussion concerning this topic. In order to be able to compare any given topic between online sub-groups it is therefore required to have a control group. To facilitate this, a 'general' news source is required to compare with a subject that falls into a certain focus group. It can, however, also be expected that on a general news website certain topics might not yield discussion whatsoever, whereas this might be actively debated on in a sub-group. With the user base being this divided, it could then also be more likely that users from different expert groups dominate discussion, if any, and therefore colour the 'general' control group in this case.

4.1.2.2 *Language*

A seemingly trivial consideration, that should be still be stated for completeness, is the language of these news websites. Due to the fact that this research, along with employing a framework for topic identification, aims to analyse the discussion on privacy within The Netherlands, it automatically limits the scope to Dutch websites only. Moreover, the topics can be assumed to come from different news sources. Either the articles and their topics are conceived by the authors or journalists of such a website, or they might originate from third-party news sources. Websites that do not follow sources of the latter case might yield a much higher amount of articles that are not date-synchronous to sites that do. This might to some extent also affect the quality of the articles as well as the amount of coverage a certain topic will receive. The quality of the articles is not necessarily a concern of the research, as the focus lies mainly on the discussion rather than the content of the articles. It is however preferred to have articles based on the same news stories as this will make comparison between for example a sub-group and a more general community possible. For example, given some news event, comparing the content of discussion between different groups. With the websites being Dutch, the ANP (Algmeneen Nederlands Persbureau) being the largest news agency in the Netherlands, forms a source for reliable news articles. The framework, however, is likely not limited to Dutch news sites only. The explicit choice for Dutch websites is therefore only due to resulting issues of a wide thesis scope, more than the method not being able to process other languages.

4.1.2.3 *Structure*

Finally, with the scope of establishing a baseline for the topic, a manually annotated list of topics reflecting the body text of the articles is required. Luckily, it is not rare for websites to supply, to some degree, labels for their content. As most websites provide a search function, for retrieval of articles within their own system, these are required to effectively handle a search query. These labels could be categories, topics or keywords that summarize and thus represent the content of the text. To simplify determining and, more importantly, checking the topic a certain document has, these labels serve an important role and requirement on which possible sources will be selected.

4.1.3 *Source Selection*

To recap, a popular news website with an active community is needed. Specifically, this website, in line with the theoretical framework, should have articles concerning privacy. In turn, it would be preferable to have a website with an 'expert' community, as well as a general news website to be able to compare with. They must be Dutch, rely predominantly on the ANP as primary (synchronous) news source, provide a comment section and labels for their articles. All these requirements summed up, it would be a logical step to take the leading general, as well as tech news websites in the Netherlands and evaluate them according to the elaborately set up requirements.

4.1.3.1 *Tweakers.net*

Tweakers.net has been around for a long time (1998). Over the years, with currently 3.5 million unique visitors and 90 million page views per month, it has become the biggest tech website of both the Netherlands and Belgium, being part of the top 20 most visited sites of the Dutch internet. At the time of writing, they report over 500, 000 registered users, with around 20, 000 online during peak time. Between 2009 and 2012 they were awarded for Website of the Year, Best News Site and Best Community. Tweakers have preserved their content well; all articles from number 1 in 1998 can be retrieved including their comments. Their community often comments on the articles posted in their news section with an average of 36 comments per article. Moreover, the articles include a 'subject' column with tags that describe the content of their news articles. Due to their news scope and community, these are considered the *experienced* in relation with digital privacy.

4.1.3.2 *Nu.nl*

Nu.nl is a general news website founded in 1999, and has published a large amount of news articles (3, 614, 791IDs at the time of writing) in categories ranging from General, Economy, Tech, Sports to things such as Lifestyle and 'Tabloid'. A separate website called NuJij provides a comment section and user registration. NuJij has, according to the website of their owner Sanoma, a unique reach of 1 million users. Nu.nl itself, however, reports between 5 and 10 million unique visitors a day. Since 2007 (last update was 2010) it has been the most visited Dutch news site. Despite the fact that it can be easily assumed that Nu.nl provides a vast amount of data, this is not true. Due to issues with ANP being its news source, the website was forced to exclude articles older than three months from their archive. Most older articles can still be retrieved for some reason; however, articles before September 2012 do not have a NuJij comment archive anymore. This greatly reduces the potential of Nu.nl as an information source, however; it still provides enough content to serve as a the *non-experienced* group for the limited time span of less than two years.

4.1.3.3 *Alternatives*

This section can be kept quite short, as the alternatives to the two news sources posted above are almost nil. Any of the possible moderately popular Dutch news sources only produce a limited amount of comments per article. 'Official' news websites that are owned by Dutch newspapers, in particular, cannot really be seen as a community and therefore generally produce less than ten comments per article. Other more subjective news websites such as Geenstijl.nl have an active community; however, the production of articles is limited to a few per week and thus might not provide enough content for the dataset. Only Tweakers.net and Nu.nl have both a sufficiently active community and a daily production of news articles following offi-

VALUE	RAW	T.NET	NU.NL	FILTERED	T.NET	NU.NL
Objects	6, 089, 559	–	–	696, 904	–	–
Size	6, 54 GB	–	–	764 MB	–	–
Articles	149, 253	91, 074	58, 179	9, 426	6, 311	3, 115
Comments	5, 940, 306	3, 256, 809	2, 683, 497	687, 474	353, 697	333, 777
Subjects	21, 100	2, 845	18, 651	2, 822	1, 221	1, 774
Users	252, 926	164, 436	88, 490	54, 987	23, 114	32, 320
Avg. Comments	39.80	35.76	46.13	72.94	56.05	107.15

Table 5: Dataset specifications providing frequencies for entities that can be found in the dataset, where RAW is the original retrieved dataset, and FILTERED that used for the experiment explained in Section 4.2. Frequencies per source per dataset are also noted.

cial news sources and therefore logically pose the sources of choice for this research, forming the *experienced* and *non-experienced* group respectively.

4.1.4 *Specifications*

For retrieval of this dataset, the scraper module from AIVB was used. A shallow introduction of this scraping process, as well as the structure of the dataset can be found in Appendix B. AIVB was written for the current research and can be used for the majority of tasks performed here. The source code can be found on GitHub[1] and is open-source under the MIT license. Finally, the specifications of the dataset can be found in Table 5. This concludes the construction of the dataset; however, some alterations have yet to be made. These will be discussed in the following section, along with the rest of the experimental procedure.

[1] https://www.github.com/fazzeh/AIVB

> *"All right," said Deep Thought. "The Answer to the Great Question...". "'Yes..!" "Of Life, the Universe and Everything..." said Deep Thought. "Yes...!" "Is..." said Deep Thought, and paused. "Yes...!" "Is..." "Yes...!!!...?" "Forty-two," said Deep Thought, with infinite majesty and calm.*
>
> – Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

4.2 EXPERIMENTAL PROCEDURE

In this section, the exact steps taken to form the retrieved dataset into a usable set for this research will be discussed, as well as the utilisation of this data as L-LDA's input for classification, the classification process itself, and the evaluation of the output.

4.2.1 *Considerations*

To recap, the entire dataset can be viewed as $\mathbb{X} = \{d_1, d_2, \ldots, d_x\}$, where d are articles with site-imposed subject[2] labels $S_d = \{s_1, s_2, \ldots, s_y\}$, comments $C_d = \{c_1, c_2, \ldots, c_i\}$ and a time signature τ_d. In turn, these comments are also provided with a time signature τ_c and will be classified having a set topics $T_c = \{t_1, t_2, \ldots, t_j\}$. The first naive assumption one could make is that given d, $\forall c \in C_d : S_d = T_c$, so that the site-imposed labels S_d are sufficient enough to accurately determine T_c for all $c \in C_d$. From there, it would be possible to simply get comment frequencies and make a time scale based on τ_c, and compare how these vary per period. In truth, however, it might be more reasonable to assume there being much noise in the dataset. A given $s \in S_d$ might be $s = \texttt{privacy}$ but $c \in C_d$ could be noise; with $t \in T_c = \texttt{cake}$. Topic Models can be employed to try and tackle this issue. By employing L-LDA, it is possible to fit a model to $\mathbb{X}$, yielding distributions over words, so that eventually all topic comments can be allocated through classification as explained in Section 3.2.3.

However, it should first be evaluated to what extent S_d are sufficient as 'ground truth', or gold standard labels for classification on a given $d \in \mathbb{X}$. Say that an unsupervised version of LDA is fitted to some amount of $d \in \mathbb{X}$, it cannot be assumed that these topics, even when setting an equal amount of topics, are similar to those when employing L-LDA to the exact same set. Similarly, it cannot be assumed that the processes of training L-LDA on some training subset $\mathbb{A} \subset \mathbb{X}$, will yield high precision on test subset $\mathbb{B} \subset \mathbb{D}$ (so that $\mathbb{B} \cap \mathbb{A} = \emptyset$), as the topics in $\mathbb{A}$ might not capture all topics in the unseen data of $\mathbb{B}$. Therefore, it is an essential first step to evaluate how the L-LDA

[2] Note that when talking about a subject s, topic labels from the website are being referred to, as opposed to a topic t as LDA uses (a probability distribution over words), which will accordingly be referred to as topics.

	T_c	
S_d	privacy	non-privacy
privacy	TP	FN
non-privacy	FP	TN

Table 6: Confusion Matrix illustrating classification errors. Given a S_d in combination with T_c, these can yield True Positives, (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

performs with subjects labels forming topic constraints, so it can be assessed to what extend they can provide a ground truth.

First, it must be considered that fitting the model with S_d creates a number of likely classification errors which can be found in Table 6. The following errors are distinguished: articles with a privacy subject label, where a comment is classified as privacy (TP), or where a comment is classified as non-privacy (FN), and articles with a non-privacy subject label, where a comment is classified as privacy (FP), or as non-privacy (TN). Due to the fact that there is no actual gold standard, these 'errors' are not deemed incorrect, they are merely used to refer to different comments and will become important throughout the rest of the experiment. So as a final example, when referring to FP discussion, this is privacy-related discussion in an article that is not labelled to be on privacy. On and off-topic will be used to refer to one article; on-topic being discussion related to the article, and off-topic not related to the article.

During the training process, it is preferred for the L-LDA model to be general enough, so that it will still allocate a low eight to the FN comments. The rationale behind this is the following: given d, most $c \in C_d$ are assumed to have a low amount of off-topic information. As such, after fitting our model, it might have enough evidence that given an off-topic c, $T_c \neq S_d$. So, given that this assumption holds, the next step is then to choose a certain amount of subjects to represent the amount of possible topics well enough. There has to be a mixture of topics large enough to distinguish different topics in the entire set of $\mathbb{X}$; however, not too much as it will quickly increase computation time and, as discussed before, produce overlapping topics.

To recap, L-LDA will first be fitted to all news articles in order to determine if the article their labels are accurate enough to fit a supervised model with. If this is the case L-LDA can be fitted to the comments with some amount of these pre-defined subjects. During fitting of this model it is hypothesized that off-topic comments will be labelled differently than on-topic ones in the same article, so that FP discussion outside of articles regarding privacy becomes mappable with sufficient accuracy. Once this is all done, the model can identify both TP and FP comments that would be on the topic of privacy.

df(s)≥ **n**	|**s**|
df(s) ≥ 1	14.855
df(s) ≥ 2	5995
df(s) ≥ 3	3854
⋮	⋮
df(s) ≥ 20	597
df(s) ≥ 40	227

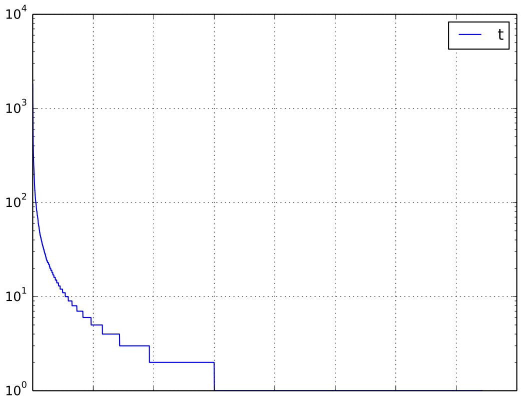


Figure 17: Left: Amount of s subjects $|s|$ with applied filters on df(s), where a certain subject t has to occur at least n times. Right: log · log scale of the subjects sorted on df(s) from the CSP of September 2012 - June 2014 where the frequency is the y-axis, their rank the x-axis, and t are the subjects.

Before this can be employed, some preprocessing and scoping of the data is necessary, however.

4.2.2 *Data Scope*

Looking at the research question, as well as what is known about $\mathbb{X}$, an exact period of time for which it is feasible to run analyses can be established. As the Nu.nl set starts from September 2012 onwards, Snowden started publishing his obtained documents in June 2013, and at the time of writing it is June 2014, the approximate of five months can be taken to describe three different periods: the pre-Snowden period (ESP), from January 2013 – May 2013, the Snowden period (ISP), from June 2013 – October 2013 and the post-Snowden period (OSP), from November 2013 – April 2014. This period scope reduces the total amount of articles from $149,253$ to $39,842$, and the total amount of comments from $5,940,306$ to $3,262,063$ in the combined period (CSP), as can also be observed in Table 5 in Section 4.1.4. From here, some amount of subject labels can be determined to construct the final dataset $\mathbb{Z}$.

Seeing as the total amount of different subject labels is 14.855 for the CSP, further analysis of the amount of documents each subject s is associated with, the df(s), is required to determine how many of these are useful for constructing $\mathbb{Z}$. As can be deduced from Table 7 the frequent subjects are very frequent, and become exponentially less frequent per subject. This results in a long tail of subjects with df(s) $= 1$, as can be seen in Figure 17. However, filtering out each s where df(s) ≤ 3 already results in a decrease to $|s| = 3854$. This choice

s	df	s	df	s	df
Smartphones	1813	Google	1222	Games	1202
Politiek en recht	1112	Microsoft	1050	Apple	1043
Samsung	918	Tablets	676	Beheer en bevei. . .	648
Syrië	595	Websites en com. . .	565	Privacy	508
Wetenschap	473	Bedrijfsnieuws	469	Sony	460
Mobiele besturings. . .	441	Android	433	Internet	400
Consoles	392	Facebook	385	LG	349
Nokia	344	`None`	342	Besturingssyte. . .	325
Laptops	325	Oekraïne	319	Mobiele netwer. . .	308
Onrust Midden-Oo. . .	304	HTC	286	Processors	265
Televisies	263	Rusland	260	Nederland	257
`empty`	254	Windows 8	249	Egypte	244
Intel	244	Verenigde staten	232	China	227
Xbox	221	Asus	218	NSA	217
Twitter	207	Verenigde Staten	198	iPhone	198
Galaxy	198	Geruchten	196	Internettoegang	194
Gameontwikkeling	194	België	191	Lumia	189
PRISM	187	Formatie	185	KPN	177

Table 7: Top 54 topics from the CSP.

is not trivial, as it has to be weighed if either quantity or quality is preferred for the input of the model; more topics means a larger mixture of possible $c \in \mathbb{Z}$ to identify. However, it also implies that subject labels with a low df(s) are likely not to provide enough content to build solid topics with. To avoid the low probability topics yielded by these, an initial aim will be made to consider at least twenty associated documents required to form a feasible subject. It is however also the case that a given d is preferred to be assigned a multitude of $s \in S_d$, in order to avoid L-LDA being a simple Bayesian Model. If the subject filter results in only one remaining subject, the whole document will be dropped. Bear in mind, however, that the primary objective is to identify comments that have a topic similar to that of privacy, rather than classifying a topic for every possible document. Accordingly, after applying a cut-off from df(s) < 20, any date before 01/01/2013, and the remaining $S_d > 1$, the total amount of articles for $\mathbb{Z}$ is reduced from $39,842$ to $9,426$, and the comments from $3,262,063$ to $687.,474$.

4.2.3 *Preprocessing*

Now that the scope is defined, it is important to first take into regard which tools will be chosen for applying any necessary preprocessing steps, as well as training, classification, and evaluation for L-LDA. There is a fair a mount of packages that implement LDA in a variety of languages[3]; however, L-LDA is seemingly supported only by the Stanford Topic Modelling Toolbox[4] (STMT), which is accordingly the module of choice. This toolbox additionally includes the ScalaNLP package[5], specifically Epic[6], that can be used for tokenization and filtering, which will be explained later on. Moreover, it supports some general preprocessing steps, which were used in combination with custom scripts from AIVB[7].

Before L-LDA is applied to the scoped dataset, some amount of preprocessing has to be done aside from the previously mentioned filtering steps. In order to provide a rough indication on how the model reacts to the input data, a qualitative analysis was performed on the output of an L-LDA model over-fitted to all articles. Some of the observations included the fact that titles and introductions often included important terms relating to the topic. While it is possible to analyse title, introduction and text respectively, here it was chosen to merge these sections for each instance, as it is preferred to have as much content as possible. This input was then tokenized with the SimpleEnglishTokenizer from the Epic package. Granted, it is a simple whitespace tokenizer and will therefore not be able to distinguish compound words, or possessive forms for example; however, it will suffice for this task. Much of the HTML clutter and other remnants from the scraping process are removed with the AIVB preprocessing module, whereas the tokenizer makes sure that only words and numbers will be regarded by the model. For Dutch, it is also possible to exclude words of length one. After tokenization documents with less than five terms are excluded, after which term frequencies are counted. Accordingly the top 150 terms, being regarded as stop words[8], are removed along with terms that occur in less than four documents.

[3] http://www.cs.princeton.edu/~blei/topicmodeling.html

[4] http://nlp.stanford.edu/software/tmt/tmt-0.3/

[5] http://www.scalanlp.org/

[6] https://github.com/dlwh/epic

[7] https://github.com/fazzeh/AIVB

[8] Although tf · idf would be likely to get rid of these, it is implemented as a standard function in STMT.

d_n	t_1	t_2	t_3	t_4	t_5	t_n
5203	.00014	.0002	$.3e-05$	.00012	.00019	$\cdots$
2688	.00012	.27732	$.3e-05$	$.4e-05$	.00016	$\cdots$
2283	$.6e-05$	.36367	.00018	$.7e-05$	$.5e-05$	$\cdots$
5397	$.4e-05$	$.8e-05$	.00029	.00656	.02684	$\cdots$
1575	$.6e-05$	.03443	$.7e-05$	.00016	$.8e-05$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 8: A sample of a document · topic matrix as outputted by the Stanford Topic Modelling Toolbox.

4.2.4 *Training & Inference*

Once the data is filtered and preprocessed, L-LDA can be trained with the eventual input. The STMT implementation requires a list of documents and their respective labels, here provided through the subject labels. It was set to train L-LDA utilizing Collapsed Variational Bayesian inference with an approximation regarding zero-order information only (CVB0) (Asuncion, 2010; Asuncion, Welling, Smyth, & Teh, 2009). An explanation on parameter optimization through inference, and the computational advantages over Gibbs Sampling and Variational Bayes which CVB0 offers can be found in Sato and Nakagawa (2012). The maximum amount of epochs to run the model was set so that the model would converge with certainty, as will be discussed further on. When a model is trained, it can be utilized as a regular LDA model to infer new tags on a test set using CVB0. The STMT package can output a multitude of output data; however, here it was set to only output a matrix with document · topic distributions (DTD) and topic usages (TU) which can be used to compare topic frequencies among different slices of data, for example by time or source. With the DTD, of which a sample is shown in Table 8, it is possible to sort the topic probabilities assigned to each document. Through this, the distribution becomes a ranked list of topics that can be used to evaluate against the original subject labels.

4.2.5 *Evaluation*

There are two steps of evaluation required before the model can be applied in the classification task of labelling the output: evaluation of the learning process, and that of the inference process. As mentioned before, the learning process is primarily optimized through CVB0, and by using a toolbox it can be assumed that to some amount it is an ironclad implementation of a Topic Model. Still, while training the model, attention has to be paid to how much the model is changing the probability weights during optimization, especially due to the fact that the STMT requires manual declaration of the amount of epochs, or iterations that CVB0 will make. By plotting the mean probability over the total amount of topics, per snapshot of 100 epochs, this optimization process can be visualized. Once the line flattens out, convergence has been reached.

More importantly, however, is the evaluation of the inference performance by the model. In order to achieve this, there exist numerous metrics that are commonly used in information retrieval (Manning et al., 2008), from which some will be utilized here. In order to be able to employ these, however, a gold standard, or ground truth is required. Since the task here is to infer a mixture of topics onto a presented document, the subject labels from dataset $\mathbb{Z}$ can be used to indicate which ones are correct, or relevant in information retrieval terms. Given some document d with subject labels S, the model infers a range of topics T with some probability. Then, for example, from T, the topics can be sorted by probability and only the first 10 displayed. Giving:

$$S = (\texttt{Lenovo}, \texttt{Tablets})$$

$$T = (\texttt{Tablets}, \texttt{Laptops}, \texttt{Lenovo}, \texttt{Ifa\_2013}, \texttt{IPhone\_5C}, \texttt{4g},$$

$$\texttt{Ascend}, \texttt{Smartphones}, \texttt{ces2014}, \texttt{IBM})$$

Most frequently used in information retrieval are the precision and recall metrics, where Precision (P) is the fraction of retrieved topic labels that are relevant, therefore:

$$\text{Precision} = \frac{|\{\text{relevant topics}\} \cap \{\text{retrieved topics}\}|}{|\{\text{retrieved topics}\}|} \tag{10}$$

And Recall(R) the fraction of relevant topics that are retrieved:

$$\text{Precision} = \frac{|\{\text{relevant topics}\} \cap \{\text{retrieved topics}\}|}{|\{\text{relevant topics}\}|} \tag{11}$$

By only considering items up until some n, a measure is called precision at n, or $P@n$ can be performed; say that there frequently are only three subject labels, it might be beneficial to only get P and R for $n = 3$. Nevertheless, seeing as one can simply get a high R scores for these measures by just returning all possible topics, a fair measure to weigh P and R with a harmonic mean, called F_1 or the F measure, is often used. Formally:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{12}$$

These metrics might perform well for single label evaluation (notice that due to the binary conversion, the multi-label information is lost); however, as Sorower (2010) argues, there are no levels of correctness for the set of labels that is the prediction of the model. Godbole and Sarawagi (2004) sought to overcome this issue in multi-label classification; by averaging P, R and F_1 over all instances, using the ground truth labels Y_i and inferred labels Z_i. Formally:

$$P = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|} \tag{13}$$

$$R = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|} \tag{14}$$

and

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \tag{15}$$

These metrics are effective when interested in an unordered set of labels. However, as LDA allocates clear weights to the inferred labels, a fair evaluation should take the order of these labels into account. Average Precision (AP) is such a metric. For each relevant topic label, it computes the proportion of relevant labels ranked prior to said label, finally averaging over all relevant topic labels. Formally:

$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i^l|}\sum_{\lambda \in Y_i^l}\frac{|\{\lambda' \in Y_i^l : r_i(\lambda' \leq r_i(\lambda))\}|}{r_i(\lambda)} \tag{16}$$

Where $r_i(\lambda)$ is the predicted rank of topic label λ for some document d_i. AP is widely used in the MAP metric, where AP is averaged over the total amount of documents D. Formally:

$$MAP = \frac{\sum_{i=1}^{Y} AP(i)}{Y} \tag{17}$$

In the experiment, MAP will be used to give a general indication of performance for the model on different slices of the data. In addition, MAP can also be used to see how well the model performs within any slice of data specifically for documents with a privacy-related topic label. This selective MAP metric will use the labels `Privacy`, `NSA`, `PRISM`, and `Edward Snowden`.

SLICE	TRAINING	INFERENCE	SOURCE	TENF
AAA	articles	articles	both	+
AAN	articles	articles	nu.nl	+
AAT	articles	articles	t.net	+
CCA	comments	comments	both	+
CCN	comments	comments	nu.nl	+
CCT	comments	comments	t.net	+
ACA	articles	comments	both	-
CAA	comments	articles	both	-
LLA	both	both	both	+

Table 9: Different data slices, their abbreviations used for reference, training and inference sets, the source that was selected on, and whether or not 10-fold (tenf) cross-validation was used.

4.2.6 *Slicing & Extraction*

In this experiment, a variety data slices from $\mathbb{Z}$ will be used as training and inference sets (Table 9) for the LDA model. A given data slice was trained through the commonly used 10-fold cross-validation if indicated. This implies that from fold 1 to 10, a model is trained on 2 to 10 and tested on 1, another model is trained on 1 plus 3 to 10 and tested on 2, and so forth. These are then all evaluated and the metrics averaged over the amount of folds. This way, the entire dataset has been 'unseen data' at some point in the rotation, so that the model never trains on instances it will have to classify during the inference process. Through this method, a realistic measure of performance is created. The slices that were not 10-folded trained on a different set than these had to infer, and therefore the inference set is unseen data.

Once the model is evaluated, and the preferred combination of slices yielding the best classification performance on comments is chosen in Experiment I, the result is then a set with classified comments coupled with some amount of topics, which can in turn be used to do an actual quantitative analysis of the privacy discussion in Experiment II. For some scale of time, say days or months, it can then for example be measured if the average amount of discussion has increased within these periods, in order to test the first hypothesis. It is specifically interesting to look at comments that have been tagged by LDA as being on the `privacy` topic, though the subject labels do not correspond with this. Ergo, these can be interpreted as being FP. Furthermore, recall that three time periods were chosen in order to compare the discussion developments and the effect the NSA disclosures had: Pre-Snowden (ESP), Snowden (ISP) and Post-Snowden (OSP). These labels can be used to test the second hypothesis for this experiment. Moreover, to test the last hypothesis the different sources, Nu.nl and Tweakers.net have to labelled and compared. To recap, for each entry five parameters have been determined: Source (Nu.nl or Tweakers.net), Period, (ESP, SP, or OSP), Scope (TP or FP), Week (coded from 1 to 22 per period), and Comment Frequency.

Finally, the Article Frequency is added so that the Comment Frequency can be normalized to obtain a reliable Weighted Comment Frequency. The Article Frequencies are extracted by counting how many unique articles were associated with TP and FP comments respectively, given a certain Week number. Say that for five articles one is on the topic of privacy, this might yield an Article Frequency of 1 and Comment Frequency of 50 in the TP scope, while four non-privacy articles might have an Article Frequency of 4 with only a Comment Frequency of 6 in the FP scope. The Article Frequencies per week within a Scope, within a Source, within a Period, are weighted by the sum of their Period's Article Frequencies and the sum of their Source's Article Frequencies. This normalizes the effect that a given Period or Source might have on the Comment Frequencies. A source might have more comments just because there are more articles in general for that source, and a period might also have more comments just because the coverage for that period on the topic of privacy was higher. By normalizing to an Weighted Comment Frequency, these unwanted effects are compensated for.

Looking at the variables, the Weeks could be used for time series based analysis; however, their order is not important when placed under the Period label. Recall that before, it was argued that in order to answer the three hypotheses, the effects of three variables on the data have to be tested: Source, Spread and Period. Accordingly, these can then easily be divided in a hierarchy with this order. For the two Sources, differences can be measured between Spread, between Periods. This mixture of between-group and repeated-measures can be tested in a mixed design, which will be performed in **R**[9].

[9] http://cran.r-project.org/bin/windows/base/

"You can't tell if a machine has gotten smarter or if you've just lowered your own standards of intelligence to such a degree that the machine seems smart."

– Jaron Lanier, *You Are Not a Gadget*

5 EXPERIMENT I: MODEL EVALUATION

In this chapter the results of Experiment I, the model evaluation, will be discussed. The primary focus lies on defining the most useful combination of data slices to use in Experiment II, as well as interpreting the results in the discussion in order to form a preview on to what extent the model will be suitable for a dependable qualitative analysis.

5.1 RESULTS

The first step as indicated in the experimental procedure is to assess if the models have enough time to optimize through CVB0. As can be observed in Figure 18, the convergence process of CVB0 already reaches convergence after 300 epochs; the mean of the allocated probabilities over topics reaches a steady line. Looking at Figure 19, heavy fluctuations can be seen at the start of the training process. However, as the mean probability differences are much lower than in for those in Figures 18 and 20, small amounts of changes in this mean strongly influence the plotted line. It can thus be concluded that a relatively

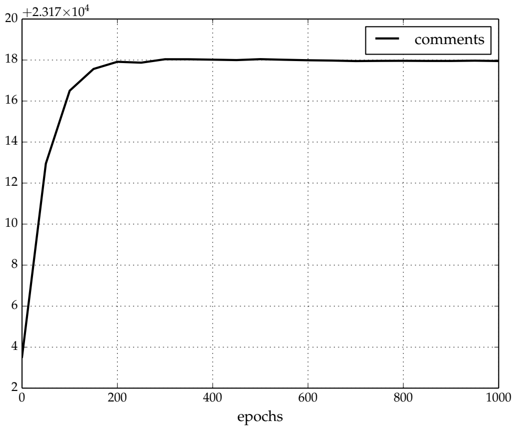


Figure 18: Model convergence while training L-LDA on CCA. The x-axis displays the mean topic probability, the y-axis the amount of iterations, or epochs.

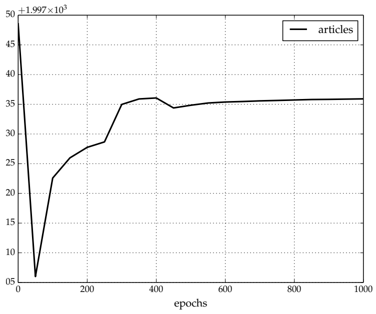


Figure 19: AAA convergence.

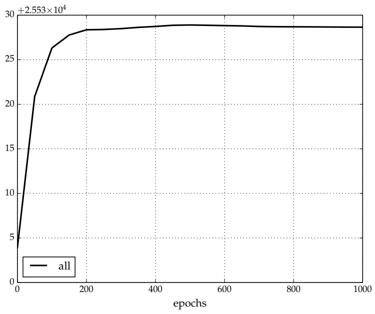


Figure 20: LLA convergence.

small amount of input almost directly results in a converged model. Given this, the combined input is expected not to train much longer, which is confirmed by Figure 20, reaching convergence around 350 epochs. So, independently of the size of the input set, the model always converges before it hits the standard setting of 1000 epochs. The training time of the models might therefore even be reduced by half through lowering the amount of epochs. Either way, it can be stated with confidence that the models have been given enough training time for CVB0 to be optimized.

The second step is to evaluate the performance of the trained models during the inference process. Table 10 displays the acquired MAP scores for the different slices of data as declared in the previous chapter. Three observations can be drawn from the part of the table that deals with inferencing topics for the comments: the scores are low, as was to be expected due to absense of a solid gold standard. Moreover, MAP decreases per n, implying that the model performs better when it has to infer only one topic, rather than more. However, at $n = 5$ the MAP score increases, at times even higher than the $n = 2$ score. This, in turn, implies that the relevant topics are commonly at the end with a high n, which indicates that a lower n cuts off these topics. Particular attention should be paid to CCT scoring higher than CCN, and CCN only scoring slightly higher than CCA on $MAP(2)$. This gives evidence of the fact that combining both sources worsens the classification success.

Another observation that can be made when looking specifically at how the inferred topics for the comments perform against the articles, is that due to a solid gold standard, the model performs considerably better at classifying the articles than it does at comments. Still, a similar pattern of score decrease per n can be observed; however, for the articles AAN seems to perform better in general than AAT. Hence, some information can be deduced about the performance between these input sets: the model is more successful in classifying articles on Nu.nl than on Tweakers.net, and similarly better in classifying comments on Tweakers.net. Overall, training on the sources separately yields higher scores than when they are combined, both

slice	$MAP(1)$	$MAP(2)$	$MAP(3)$	$MAP(5)$
AAA	.688 (.616)	.535 (.476)	.511 (.442)	.532 (.478)
AAN	.713 (.796)	.537 (.623)	.551 (.630)	.580 (.660)
AAT	.703 (.629)	.557 (.496)	.519 (.470)	.537 (.496)
CCA	.336 (.256)	.214 (.161)	.198 (.147)	.205 (.152)
CCN	**.334** (.268)	**.215** (.166)	**.210** (.160)	**.222** (.171)
CCT	**.390** (.389)	**.256** (.255)	**.221** (.218)	**.223** (.220)
ACA	.113 (.091)	.090 (.072)	.090 (.071)	.100 (.079)
CAA	**.730** (.573)	**.548** (.422)	**.517** (.393)	**.530** (.409)
LLA	.344 (.263)	.221 (.166)	.205 (.151)	.213 (.157)

Table 10: MAP scores for each data slice. The selective MAP score, which only evaluated on articles that were privacy-related according to the gold standard, is noted behind parentheses.

for comments and articles. Moreover, combining both comments and articles in the LLA slice does not result in a notably better score when inferring on that set. It can therefore be concluded that by combining the sets, the model performs worse on inferring topics for the articles and equal or worse on comments in that particular slice. Finally, training on articles and inferring on comments in the ACA slice yielded the lowest scores in any data slice.

In addition to the MAP scores for all articles and comments, a selective MAP score was performed on each data slice, only looking at articles and comments that had a privacy-related subject label. The scores for this measure are also shown in Table 10 behind parentheses. It can be observed that, looking at the inferred comment topics in the CC section, for Tweakers.net the scores are almost equal to the non-selective MAP score, while for Nu.nl these are noticeably lower in comparison. The opposite is true when looking at the inferred article topics in the AA section; for Nu.nl the selective MAP score is even higher than the non-selective. This is the same clear distinction in scores between sources on both articles and comments that could also be observed for the non-selective MAP score.

Finally, the results for training the model on comments and testing on the articles, thus again with a more solid gold standard and evaluation score than when testing on comments, can be found in the CAA row of Table 10. As can be seen, the performance for this model is higher than when both training and inferring topics for the articles, across al n except for $n = 5$. However, it performs worse at classifying the articles that are on the topic of privacy in the non-selective MAP score.

"One day you'll walk the world and keep in mind,
the heart you've been given in winter time."

– Gojira, *Born in Winter*

5.2 DISCUSSION

So far, the results of training L-LDA on the collected and sliced dataset have been investigated. Before going into the part of statistical analysis on the classified instances, however, it is important to assess the implications the results obtained thus far hold, and the influence these might have on the next experiment. First of all, it can be concluded that L-LDA handles the amount of labels and data well, as it does not need an excessive amount of epochs to converge. Despite the fact that testing differently preprocessed inputs, a variety in label amounts, and different sampling methods or other model optimization tasks were beyond the scope of this thesis, at least it can be stated that the current model's results are reliable. In this sense, the results are to be interpreted bearing in mind that these are not ironclad results; they display how well the model performs given the previously made choices. However, it is our belief that these choices are well grounded throughout the experimental design, and the results can therefore be interpreted, with confidence, in accordance with the framework of this thesis.

Given this, it is possible to interpret the evaluation metrics more elaborately. As seen before, the model is quite successful in classifying articles. Seeing as the model is provided with labels that directly linked to the subjects of the articles provided by the website, this is arguably not too surprising. However, it is still an important result; low scores here would have implied that the subject labels were insufficient even for labelling the articles correctly. Apart from this being due to the fact that L-LDA could not have worked as a model to begin with, some other explanations that yield lower scores can be given. Websites could for example have provided numerous labels that were not related to the content of the article at all, merely for the article to show up in more searches. Concrete examples of this can be found when delving through the classified instances: an article on `Motorola` that is also tagged with `Google`, as Google owns Motorola since 2011. Looking at the topic distribution in the learned model, there logically is no evidence supporting a connection between Google and Motorola, as this would have required articles from 2011. A lower score could therefore be seen as analogous to noise in the subject labels.

In this sense, it can be evaluated whether a source provides accurate subject labels, and to what extent these can be learned. Accordingly, it can be stated, looking at the AA section of Table 10, that Tweakers.net provides less accurate labels than Nu.nl, especially for privacy-related topics. The fact that combining the two sources yields an even lower score is evidence of too much total, or similar topics that the model has to chose from, which results in overlapping topics that might still be close to the gold standard. This could explain why a label such as privacy, which is likely to overlap between sources, scores worse than the lowest score between sources in AAA.

To interpret the scores in the CC, recall that the article subject labels were used to evaluate the successfulness of classifying the comments. So, even if they are on a completely different topic, the comments are still labelled according to the articles. Accordingly, it was expected a priori that the comment evaluation scores would be low, as the gold standard might be incorrect. However, through assuming that there are not that many FP comments, they might still be reliable to some extent. In comparison with the AA section, the CC MAP scores seem quite low. However, bear in mind the complexity of the task, and the existing evaluation handicap. Another interesting observation was the difference in MAP score between Tweakers.net and Nu.nl. Due to the article imposed gold standard labels, a higher score for a given source would imply that the comments are more frequently off-topic on average. Given this, it can be stated that comments on Nu.nl are more frequently off-topic than those on Tweakers.net. In turn, the notably lower score for privacy-related articles for Nu.nl are evidence of this topic being more prone to off-topic commenting for this particular source.

Furthermore, looking at the bottom part of Table 10, for ACA in particular we can see clear evidence of how the model fails to classify the comments when training on the articles. This is a strong indication of how the amount of content in the comments differs greatly from that of the articles. Note however, that leaving comments out of the training process also greatly increases the chances of detecting off-topic comments; by only offering a much smaller amount of information it can be argued that despite the topics being sparse, they are are not fitted with topics that might be off-topic, but also persistently used related topics in off-topic discussion. Privacy and security might be an example of repeatedly co-occurring topics within a discussion. This is mere speculation, however, as more accurate labels would be needed for comments to be able to test this. Secondly, the combination of articles and comments in LLA has not noticeably improved the MAP score if it compared to CCN and CCT. At first it might appear to be better than CCN; however, the selective MAP score indicates that it actually performs worse for the task in this thesis.

It was discussed before how the evaluation for comments was predicted to give a pessimistic results due to the fact that the gold standard is inaccurate for off-topic comments. However, as was hypothesized while discussing the experimental procedure, L-LDA is likely to allocate low probabilities to off-topic noise in its topics, and accordingly, still form a solid model even with partly inaccurate labels. This is exactly what the CAA part of Table 10 confirms. When training on the comments, the articles classified articles yield a higher MAP score than when training on the articles themselves. It should be noted, though, that these scores cannot be directly compared as being better or worse, as the training/testing set combinations are different. From this, however, the assumptions made in the previous paragraph can be rejected; comments do not add noise to the model; instead, they are likely to provide more content from which L-LDA can deduce topics that are richer in terms of word variety and can therefore more successfully classify the articles. This bodes well for the classification of comments; although the evaluation measure is low, it can still be argued that performance is expected to be in line with that of the article classification task. As such, the classified output from the model is argued[1] to be a reliable enough reflection of the privacy topic.

Conclusively, the MAP evaluation can help to determine the optimal n amount of topic labels to be used to determine the topic labels for the set used in Experiment II, with the best balance between retrieved labels and model performance. Recall that this set will only compare Weighted Comment Frequencies, and therefore try to determine the amount of TP and FP comments, as was discussed in Section 4.2.6. The MAP measure seemed to fare best with $n = 2$; however, it could be argued that only classifying an article with two labels might be too strict. As such, both $n = 2$ and $n = 3$ will be considered in the first part of the statistical analysis.

[1] It must be kept in mind that there is no baseline to compare the MAP performance with, and therefore the performance cannot be compared objectively. If Experiment II can be conducted with some success, however, it can be argued that the model is sufficiently useful for this research.

> *"The reason that you're in the mess that you find*
> *yourself in is because you have a curious mind."*
>
> – Andy Field, *Discovering Statistics Using R*

6 EXPERIMENT II: ANALYSIS

Having looked at the implications Experiment I posed for the statistical analysis in this chapter, the focus will now lie on reporting and discussing tests on the final weighted dataset.

6.1 RESULTS

Before reporting the results for the statistical analyses on the dataset, more qualitative observations can be drawn from the visualization of its figures. Two initial plots are displayed in Figure 21, where raw frequencies for both articles and comments are displayed per source per month. At the very start of measuring, TP Tweakers.net comments surpass that of Nu.nl. Then, from March onwards, an increase in Nu.nl discussion can be observed. This is the case for Tweakers.net as well; however to a lesser extent. After this peak, the discussion on Tweakers.net still rises while Nu.nl diminishes. Finally, around December both comments and articles start to dwindle for both sources. Note, however, that the amount of comments in particular is higher for both Nu.nl and Tweakers.net than it was at the start of measuring. A final important observation from this plot is the fact that even in raw frequencies, Nu.nl has a sparse first period. The TP plot indicates that these fluctuations are less pronounced for FP comments. More-

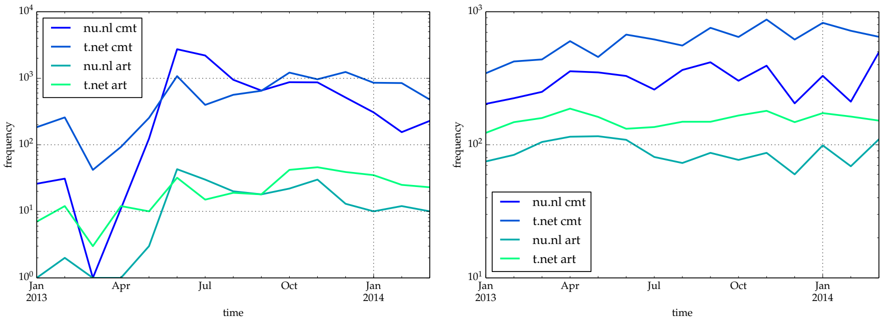


Figure 21: Frequencies of TP (left) and FP (right) comments and articles over time for each source on a log scale.

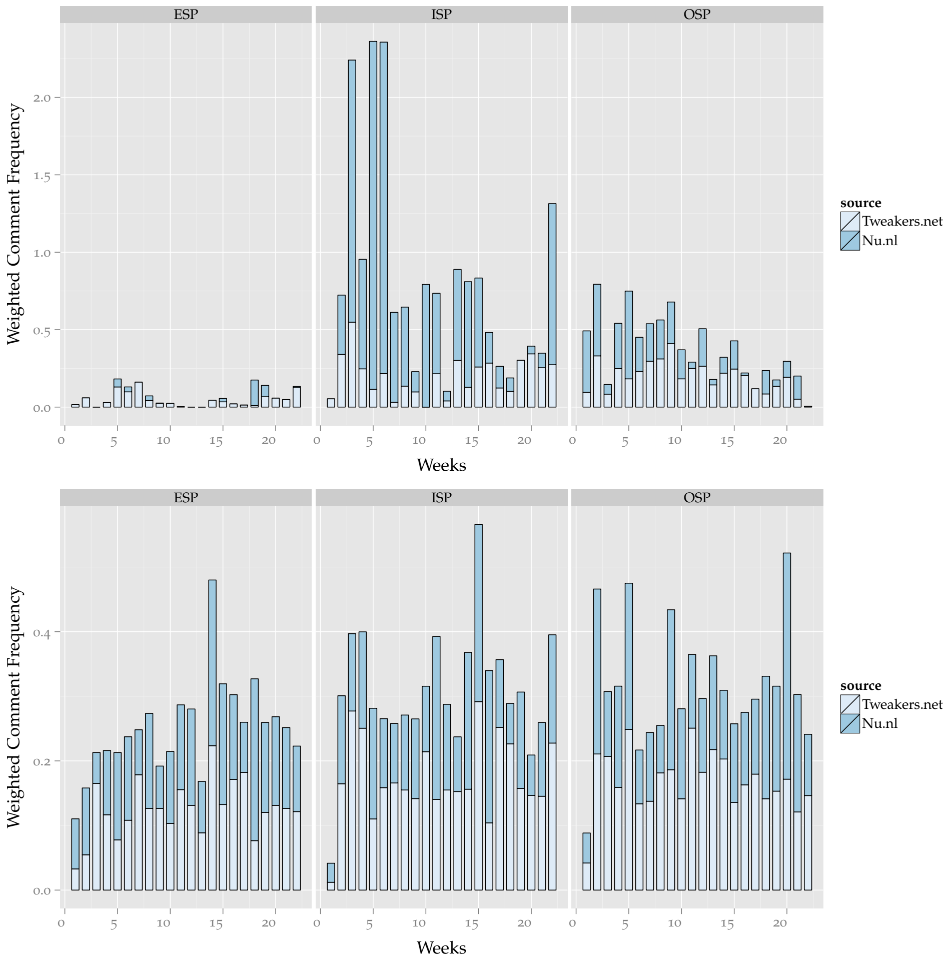


Figure 22: Visualization of the final weighted dataset for $n = 2$, with frequencies per week, split by period for both TP (top) as FP (bottom) comments. Descriptives are given in Table 11.

over, the number of articles these FP comments have been posted in are higher as well.

If these numbers are weighted, however, it can be expected that especially during the ISP, Nu.nl will have a higher average than Tweakers.net due to there being more comments for fewer articles. Having plotted the final weighted dataset for $n = 2$ in Figure 22, a more detailed representation of the data is given. First looking at the top plot, the strong differences between periods for Nu.nl are clearly visible, as well as the expected differences between frequencies between sources. As for Tweakers.net, the differences are not that strong between ISP and OSP. The bottom plot shows that for FP comments, both sources seem to increase in frequency over time. Finally, for both ISP and OSP, in week 1 only a small amount of comments seems to be classified.

Having drawn some initial information from plotting the dataset, the statistical analysis was performed afterwards. Recall that in the experimental procedure it was not yet decided whether $MAP(2)$ or

	$n = 2$			$n = 3$		
slice	mean	var	st.dev	mean	var	st.dev
FP_T_ESP	0.125	0.002	0.044	0.157	0.003	0.050
FP_T_ISP	0.173	0.004	0.064	0.191	0.004	0.064
FP_T_OSP	0.169	0.002	0.046	0.201	0.003	0.051
FP_N_ESP	0.125	0.003	0.053	0.168	0.004	0.065
FP_N_ISP	0.136	0.004	0.063	0.177	0.006	0.079
FP_N_OSP	0.147	0.005	0.071	0.202	0.009	0.096
TP_T_ESP	0.046	0.002	0.046	0.042	0.002	0.041
TP_T_ISP	0.201	0.017	0.131	0.179	0.013	0.115
TP_T_OSP	0.195	0.010	0.099	0.189	0.009	0.095
TP_N_ESP	0.017	0.002	0.039	0.017	0.001	0.046
TP_N_ISP	0.601	0.429	0.655	0.599	0.423	0.651
TP_N_OSP	0.182	0.023	0.153	0.190	0.025	0.159

Table 11: Descriptives of the weighted data for $n = 2$ and $n = 3$ per slice, with from left to right TP/FP, Nu.nl/Tweakers.net and the period indicator per slice. Reported are the mean, variance (var) and standard deviation (st.dev).

$MAP(3)$ was to be chosen for the final dataset. By looking at both the dataset its descriptives and the associated boxplots, the differences in performance can be assessed. Note that quality has already been measured through MAP; however, a lower n could result in not having enough classified data for analysis.

So, looking at Table 11 a number of observations can be made. First, the means tend to increase for FP comments between $n = 2$ and $n = 3$, and to decrease TP ones. Accordingly, for the TP comments, the variance diminishes through a higher n, decreasing for the FP comments. Looking at the means between periods, and comparing n herewith, for FP Nu.nl and Tweakers TP respectively, ISP and ISP trade-off in the highest average when increasing n.

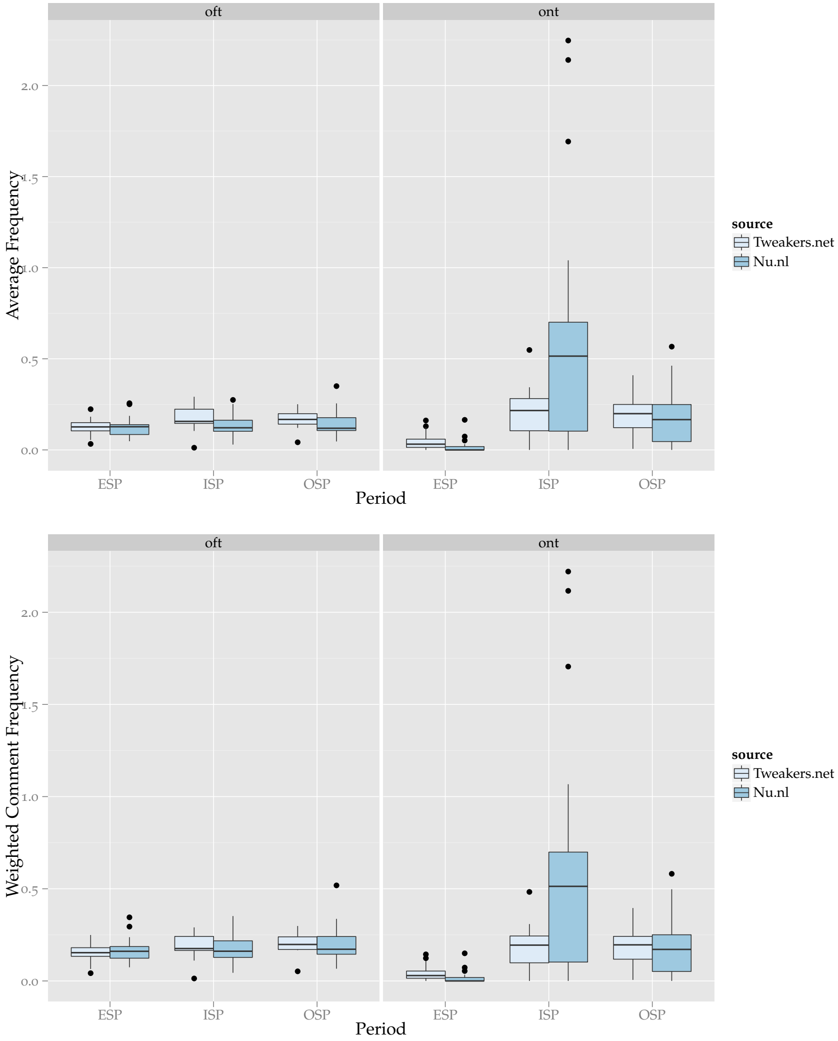


Figure 23: Boxplot for $n = 2$ (left) and $n = 3$ (right).

Looking at Figure 23 gives a clear view on the subtle differences. It must be noted, however, that $n = 2$ does add extra outliers to the FP comments. Hence, the differences between $n = 2$ and $n = 3$ do not seem to outweigh the difference in MAP score. The statistical analysis was therefore decided to be conducted on the $n = 2$ set.

However, looking at the numerous other outlier dots, it can be concluded that the distribution must be skewed. Plotting the distribution in Figure 24 at the left hand side gives clear evidence for this skew. A Levene's test indicated that the variances in the data are significantly different with $F(2, 261) = 18.04$, $p < .001$ between sources and therefore the homogeneity of variance assumption has been violated. Even when transforming the data using a square root and log transformation (Figure 24, right side), the test is still significant with $F(2, 261) = 6.28$, $p = .002$. This ruled out the possibility of doing a Mixed Model ANOVA as a parametric test.

check tenses to be present throughout results

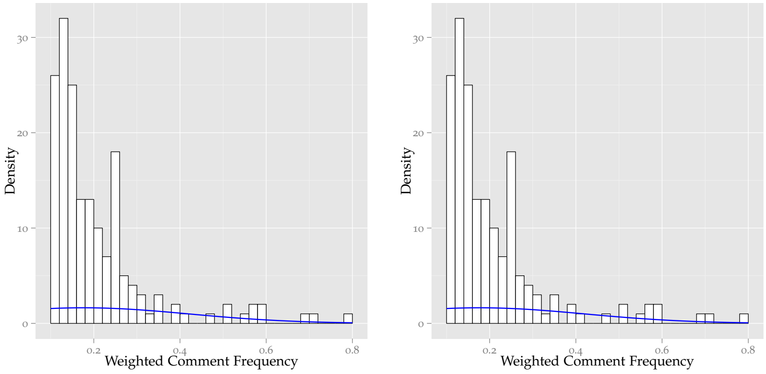


Figure 24: Distributions of comment frequency; the original data left and the transformed data to the right.

Instead, Wilcox' robust ANOVA for mixed designs was used. By doing so, a three-way scope · source · period analysis could not be performed, however, as there simply is no such implementation. The scope is therefore disconnected from the interaction analysis and will be reported individually. Accordingly, source is used as factor A and period as factor B. It was decided to perform the M-measures and bootstrap variant of the `WRS` package[1] as it empirically determines how much of the mean should be trimmed, rather than this having to do be done by hand in the trimmed means variant of the test. It can be reported that there is no significant main effect of source, $\hat{\psi} = -0.07$, $p = .216$, for TP comments. The main period effect is significant, however, $\hat{\psi} = -0.22$, $p < 0.01$. Subsequently there is no significant interaction effect between these, $\hat{\psi} = 0.29$, $p = .296$. For FP comments, it can be reported that there is a significant effect for the main effect of both source, $\hat{\psi} = 0.03$, $p = .004$, and period, $\hat{\psi} = -0.03$, $p = .003$, and no significant results for their interaction effect, $\hat{\psi} = -0.03$, $p = .619$.

The period effect does not give much information unless it is split up to form a two by two design for the three period combinations EI (ESP and ISP), IO (ISP and OSP), and EO (ESP and OSP). First TP comments: for EI this yields a non-significant main effect of source, $\hat{\psi} = 0.11$, $p = .153$, a non-significant effect for period, $\hat{\psi} = 0.29$, $p < .001$, and a non-significant interaction effect, $\hat{\psi} = 0.29$, $p = .064$. For IO it can be reported that there is no significant main effect for either source, $\hat{\psi} = 0.11$, $p = .165$, or period, $\hat{\psi} = 0.06$, $p = .238$, and neither was the interaction effect significant, $\hat{\psi} = 0.23$, $p = .196$. Then, for the EO test, there is again no significant source effect, $\hat{\psi} = 0.11$, $p = .212$, although the main period effect is significant, $\hat{\psi} = 0.14$, $p < .001$. Accordingly, the interaction effect is non-significant, $\hat{\psi} = 0.01$, $p = .784$.

[1] https://github.com/nicebread/WRS

scope	comb	source	period	inta.
TP	all	-	+	-
	EI	-	+	-
	IO	-	-	-
	EO	-	+	-
FP	all	+	+	-
	EI	-	+	-
	IO	+	-	-
	EO	+	+	-

Table 12: Summary of the two-way robust ANOVA test results, noting combination (comb) of periods the test outcome, for the main effect of source, period and the interaction (inta.) effect.

Second are the FP comments: for EI a non-significant source effect, $\hat{\psi} = 0.02$, $p = .122$, a significant period effect, $\hat{\psi} = 0.03$, $p = .002$, and a non-significant interaction effect, $\hat{\psi} = 0.03$, $p = .214$, can be reported. For IO a significant main effect for source is found, $\hat{\psi} = 0.04$, $p = .001$, as well as a non-significant period effect, $\hat{\psi} = 0.004$, $p = .812$, and a non-significant interaction effect, $\hat{\psi} = 0.005$, $p = .882$. Finally, the EO test yields a significant effect for period, $\hat{\psi} = 0.03$, $p = .041$, and source, $\hat{\psi} = 0.03$, $p = .032$; however, no significant interaction effect, $\hat{\psi} = 0.01$, $p = .717$.

To clarify the results, an overview of the results is given in Table 12 and the implications for each effect will be noted. The main effect of source reflects the average comment frequency being higher for a certain source, whereas the main effect of period reflects the fact that it is higher between certain periods. The interaction effect would reflect the fact that between sources, the periods would be significantly different in terms of comment frequency. In combination with Figure 25 these results can be interpreted. In the table, it can be observed that there is no period in which the TP comment frequency differs between sources, and therefore there are no differences between their sources' periods. The periods themselves, however, differ equally between TP and FP comments. For both there can be observed a significant difference between ESP, and both ISP and OSP respectively in Figure 25. There is no difference between ISP and OSP in either of the cases. Finally, for FP it can be determined that there is a significant difference between sources for IO and EO, both in the favour of Tweakers.net.

As a final test, a direct comparison between sources only regarding each period individually, between sources, for both on and FP comments, was conducted. The TP group has a non-normal distribution for ESP, $W = 0.74$, $p = .00$, and requires a non-parametric Wilcoxon signed-rank test. However, the FP group was found to be normal,

scope	period	normal	t.net	nu.nl	sig
TP	ESP	-	>	<	+
	ISP	+	<	>	+
	OSP	+	>	<	-
FP	ESP	+	>	<	-
	ISP	-	>	<	+
	OSP	+	>	<	-

Table 13: Summary of the Wilcoxon signed-rank and dependent t-tests, noting for each of the periods firstly if the distribution was normal (plus, t-test) or non-normal (minus, Wilcoxon), secondly which source was larger (>) or smaller (<) for both Tweakers.net (t.net) and Nu.nl (nu.nl). Finally it is noted if this difference was significant, where a plus indicates significance, a minus non-significance.

$W = 0.95$, $p = .060$, for which a dependent t-test will be conducted. ISP TP is normal, $W = 0.96$, $p = .182$, and FP is not, $W = 0.69$, $p < .001$. OSP is normal for both TP, $W = 0.97$, $p = .247$, and FP, $W = 0.96$, $p = .089$.

So, for ESP TP comments, average comment frequency is significantly higher for Tweakers.net (Mdn $= 0.032$) than for Nu.nl (Mdn < 0.001), $p = .002$, $r = -.46$. Subsequently, for FP comments, the difference of comment frequency between sources is not significantly lower for Tweakers.net (Mdn $= 0.126$) than for Nu.nl (Mdn $= 0.127$), $t(21) < 0.01$, $p = .999$, $r = -.02$. For ISP TP comments, the average comment frequency for Tweakers (Mdn $= 0.216$) is significantly lower than for Nu.nl (Mdn $= 0.515$), $t(21) = 2.92$, $p = .008$, $r = .54$, while for FP comments, the average comment frequency is higher for

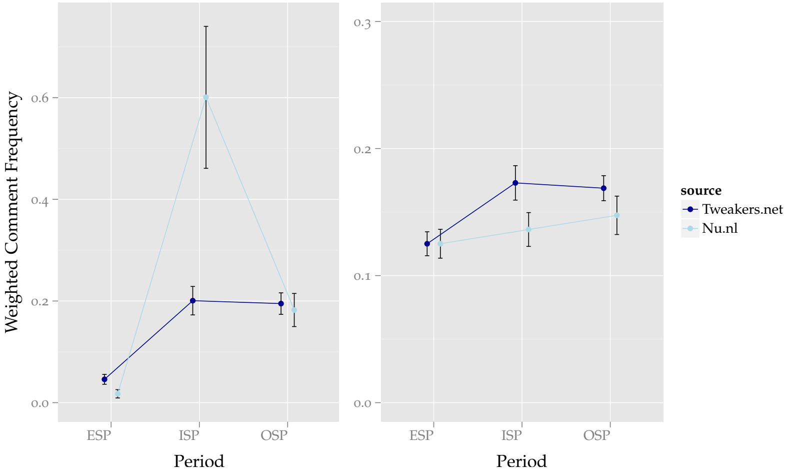


Figure 25: The mean average comment frequency as a function of their originating source and which period these were in. Left for TP, right for FP comments.

Tweakers.net (Mdn = 0.157) than Nu.nl (Mdn = 0.122), $p = .030$, $r = -.33$. Finally, for OSP TP, average frequency for Tweakers.net (Mdn = 0.199) is not significantly higher than Nu.nl (Mdn = 0.166), $t(21) = 0.41$, $p = .685$, $r = .17$, neither is there a difference for FP comments between Tweakers.net (Mdn = 0.167) and Nu.nl (Mdn = 0.119), $t(21) = 1.40$, $p = .177$, $r = .29$. These results have again been summarized in Table 13.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

– Ronald Fisher, *Sankhyā*

6.2 DISCUSSION

With the results gathered in the previous section, the hypotheses stated in Section 2.3 can be put to the test, and the research questions can be answered. The first hypothesis stated that due to the increase in media coverage, it could be expected that the discussion surrounding the topic of privacy has increased. According to the results obtained through classification of online discussions, it was found that between the five months prior to Snowden's document leaks and the five months in which this was ongoing, discussion increased significantly on both Tweakers.net and Nu.nl. This effect could be observed for discussion on the news articles on the topic of privacy, as well as that of articles covering news not related to privacy. Recall that through normalization of the dataset any influence the article frequencies per source and period might have had were corrected. Therefore, these results can be interpreted as follows: after Snowden started leaking documents, the amount of comments on articles covering this, and related news stories, increased significantly. Moreover, users on these websites also started commenting significantly more regarding this topic outside of these news stories.

More importantly, the effect was found to be persistent in the five months after the peak of Snowden's publications. This implies that an initial reveal of the NSA's actions did not only increase discussion, it also sustained this increase in the months after. This might be caused by the previously discussed societal intrusiveness of these actions (Fuchs, 2011; Westin & Blom-Cooper, 1970); one might expect that after a certain news event had its prime, discussion dwindles. However, according the results presented here, this is not the case. Granted, the longitudinal effects have not been measured, as this data is yet to be created. Moreover, it could be argued that for a very strong test of the prior discussion it would have been beneficial if Nu.nl had data available before September 2012. Alas, the latter is not the case, and it would therefore have thrown the timespan of the compared periods out of balance. The timespan used here was therefore decided to be the only possibility. Given this, it can be concluded that in line with the theory that exposure to, and awareness of, privacy risks raises suspicion, concern, and protective behaviour (Bellman et al., 2004; Buchanan & Paine, 2007; Drennan et al., 2006; Sheehan, 2002), it also directly affects online discussion and has a sustained effect given the measured period.

The second hypothesis required an assessment of two factors: the comment frequencies in the period prior to Snowden's leaks, as well as the difference between sources. In the final test of the results it was found that TP comment activity was higher for Tweakers.net, with a medium to large effect size, and in turn, that there was no difference between FP commenting whatsoever, with a very small effect size. This can be interpreted as follows: as an inexperienced group, Nu.nl users were less active on articles that were on the topic of privacy. Concurrently, this could imply that due to the fact that Tweakers.net users formed the experienced group, they might have had more to discuss in-depth about on these articles. Therefore, the hypothesis that Tweakers.net activity was thought to be more noticeably present during this period can be partially confirmed. Surprisingly, there was almost no difference for FP comments, whereupon one might argue that the latter weighs more heavily in assessing this hypothesis. The rationale behind is that the notion of awareness might be better captured by the fact that FP comments are higher; the event of the topic of privacy being triggered in a news article that is not directly related. In this sense, it can be argued that Tweakers.net only seemed to have significantly more discussion on the topic of privacy for articles that already discussed the topic. From there it might be more realistic to partially reject the hypothesis that the technically experienced group was more actively spreading discussion on the topic of privacy prior to the Snowden period.

Accordingly, the previous assumption that Tweakers.net can be deemed technically experienced and Nu.nl inexperienced might be questioned. As was stated before, however, the characteristics of these groups were determined by the target group of the individual websites. Logically, there is a chance that IT-related news on Nu.nl draws technically experienced people to comment; however, it was assumed that the majority of the target group was technically inexperienced. This assertion would require a comparison between word use, for example, to be tested empirically, which would be a different research altogether. Conclusively, it can be stated that the prior awareness of the technically experienced group only resulted in larger discussions on the topics themselves, rather than a significantly higher amount of FP discussion. Hence, this hypothesis is partially confirmed.

So far, the difference between the period prior to Snowden's leaks, compared to those during and after has been looked at independent of source. Thereafter, this prior period was compared between sources. However, the interaction between these periods and that of the sources, or groups, has not yet been considered. In line with this particular comparison, the third hypothesis stated that a stronger impact on the inexperienced group was expected as a result of the leaks. To test this, the results of both the robust ANOVA (from now on referred to as the Mixed tests), and Wilcoxon and t-test (referred to as the Mean tests) are to be interpreted. To say that the inexperienced group would be more heavily impacted by the news surrounding the NSA, the period in which these were published, the Snowden period, should be looked at specifically. As was reported in the Mean test, the users of Nu.nl commented much more TP during this period than those of Tweakers.net, as could be deduced from the high effect size. In contrast, Tweakers.net had a significantly higher FP comment frequency than Nu.nl during the same period. No interaction effect was found in the Mixed test, however, which implies that there was no significant increase between sources, between the periods prior to, and during the leaks.

An explanation for this requires considering the period after the leaks as well; here the difference between sources equalized according to the Mean test, but were still significantly more than the first period according to the Mixed test. Observing and summing up these effects yields the following final conclusion that clearly maps the effect of topic exposure: for TP commenting, there was a very strong direct increase for Nu.nl when the leaks became news, and while there was no significant increase between sources in the transition between these periods, Nu.nl discussion expanded significantly during the period of the leaks. As such, it can be confirmed that the increasing awareness and growing attention to the existing privacy discussion, which was expected to be novel to the majority of Nu.nl users, heavily affected the amount of discussion on the articles covering it. In combination with a comparable effect for Tweakers.net, it can be stated that topic exposure directly increases discussion.

For FP comments the effects are not that clear-cut. Users from Tweakers.net, where the majority of users was expected to already be familiar with the topic, saw a particular strong increase for FP comment frequencies during the leaks. According to both the Mean and the Mixed test, this increase was higher for Tweakers.net than for Nu.nl during that particular period. After, however, through a steady increase in FP comment frequency for Nu.nl, this equalized between sources. Therefore, it can be put forward that the experienced group saw a faster increase of spreading the topic to articles that were not related, and the topic exposure effect was delayed for the inexperienced group. This could be explained in twofold: due to the fact that Tweakers.net's content is solely on IT-related topics, there is an increased amount of comments that allow the discussion of this topic that is interwoven with many IT topics. This would explain the significant increase for the period during the leaks. Finally, it can be argued that the Nu.nl equalization due to a steady increase of FP comment frequency could be explained through the notion of awareness. Through increased awareness, inexperienced users might start to increasingly relate the topic of privacy to more than just the related articles, whereas Tweakers.net users could have already had this association to begin with. Either way, it can be concluded that the hypothesis can be partially confirmed; Nu.nl were affected more heavily by the initial news regarding the leaks, but took more time to increase in average FP comment frequency. Low topic awareness is therefore argued to be linked to a strong reaction on the initial news, and a delayed discussion spread.

"Vision will blind. Severance ties. Median am I. True are all lies."

– Meshuggah, *Sum*

7 CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

This research sought to tie in with the societal interwoven concept of privacy; specifically the growing attention digital privacy has gained through the years, and the radical change its notion has been subjected to after Edward Snowden revealed the mass-surveillance activities that numerous intelligence agencies are involved in. Given the fact that the Netherlands has had an extensive amount of coverage on this topic, its public debate was considered to be mostly avoided by the government. However, it was argued that there was enough reason for this discussion to already have manifested online. In turn, this put forward the novel research opportunity of assessing this discussion in light of this shift in the concept of digital privacy.

In particular, the aim was to unravel how the discussion on the topic of privacy was impacted by the news covering the leaked activities of the NSA. Through reviewing theory on the underlying factors of this privacy debate, it was established that topic exposure, and prior technical education increased suspicion regarding privacy risk. Accordingly, these formed important factors in quantifying the differences between inexperienced and experienced groups in topic exposure and correlation with the frequency of online discussions.

Assessing to what extent these effects where prevalent required a novel approach for computational analysis of online discussions. It was hypothesized that through the application of a Topic Model to discussion in online news communities, developments over time between these communities could be effectively classified and analysed. To facilitate this, a dataset was constructed containing articles and discussion of general news website Nu.nl from September 2012 onwards, and the entire collection of articles and discussion from information technology website Tweakers.net.

The first experiment conducted in this research evaluated the Labelled Latent Dirichlet Allocation (L-LDA) model on its performance as a classifier of online discussions. It was found that the model performs well on the classification of news articles with the subject labels provided by the website; however, as the same labels were used to train on the comments forming the discussions, the evaluation would always give a very pessimistic score on comment classification, and therefore expected low scores were found for this task. However, using a topic model that was trained on the comments and inferred on the articles, a reliable alternative to evaluate the task of classifying the comments could be produced. This showed that the topic distribution fitted by comments performed comparably as that of articles. As such, it could be concluded that, in turn, L-LDA proved to be a promising model for classification of online discussions on these news communities, using large amounts of small input. Moreover, it was argued that L-LDA handles the off-topic noise in these discussion well, and that the assumption of imposing the comment's parent subject labels, provided by the articles are a sufficient, albeit rough, golden standard. Through lack of comparison of the scores, these results had to be framed within the research of this thesis, and were argued to not be easily generalizable. Having established that the used Topic Model performs sufficiently on extracting topics from the dataset as well as classifying these for the posed research, the extracted data from this model was statistically analysed in order to test the hypotheses regarding the topic exposure differences, which will be repeated here.

First it was sought to be investigated to what extend the Dutch discussion on digital privacy increased. It was found that overall, discussion regarding the topic of privacy has increased for both the non-experienced group of Nu.nl, and the experienced group of Tweakers.net. Moreover, the effect was persistent ten months after Snowden's first leak, which confirmed the hypothesis that topic exposure for this particular topic directly affects online discussions. Secondly, it was to be investigated to what degree the hypothesis regarding an a priori difference in discussion due to experience held. It was hypothesized that the experienced group would be likely to already have been discussing the topic of privacy before the leaks. This was not completely confirmed, however, as Tweakers.net users only participated significantly more in discussion that was already being conducted on the articles concerning this topic. Off-topic, there was almost no difference between groups. Finally, and tying in with the thesis, it was aimed to investigate how the leaks affected both groups respectively. Here, the topic exposure effect was further confirmed through comparing the periods during, and after Snowden's publications. It was found that it had a direct effect on the discussion on the articles concerning privacy, which was strongest for the non-experienced users of Nu.nl. The effect in the period after was less for Nu.nl, although for both groups it was still significantly more than the period prior to the leaks. This reaction on articles on the topic of privacy for Nu.nl could, in contrast, be observed for Tweakers.net in an increase of spreading discussion on articles that did not concern privacy. However, here, for Nu.nl the latter increased to finally equalize with the amount of Tweakers.net, both being significantly more than the first period of measuring. Returning to the stated thesis, it can only be partly confirmed that topic exposure influences discussion for both groups differently; however it was confirmed that both groups saw a significant increase in overall discussion. As such, the hypothesized manifestation of online discussion surrounding the topic of digital privacy in the Netherlands can be confirmed. It might be even stated that regarding this increase of discussion, we really do live in a post-Snowden era.

> *"You'd be amazed how much research you can get done when you have no life whatsoever."*
>
> – Ernest Cline, *Ready Player One*

7.2 FUTURE WORK

Through the successful collection of a large dataset, and training of a reliable initial version of a topic model, it has been demonstrated that these models can be effectively employed for research on the developments within a public discussion, as well as topic exposure. Accordingly, the possibilities for extending the research in this thesis are numerous. In this section, some ideas for future work in line with this thesis will be discussed, where the focus will lie primarily on different fields of research that could extend the current research.

A first effort of extending the current work could be made by assigning a broader scope to the topical focus in particular. In contrast to focussing on privacy alone, the model in this thesis is ready to be used on any of the different topics that the dataset holds. In this regard, it could be further evaluated how the computational approaches might perform when focussing on different important topics of discussion on news websites. More specifically, it could be extended to not only analyse developments within discussion frequency over time, but also to detect exactly when a certain topic becomes more important in a certain time frame. A related implementation is already available L-LDA, which can split the distributions over topics into periods of time. From there, these probabilities could be compared in a similar fashion as in this thesis to test if they significantly increased, indicating some amount of increased importance. An additional form of research could then be conducted in analysing the interaction between topics. For example, say that the system detects the topic of privacy to have increased, it could be hypothesized that certain related topics (such as security) might also increase.

A particularly interesting approach here was left out of this thesis due to time constraints; to implement event identification which has already been successfully applied to social media (Becker, Naaman, & Gravano, 2011; Vavliakis, Symeonidis, & Mitkas, 2013). Say that at some time it is identified by the system that there is a significant, longitudinal peak in topic frequency at a certain day, or span over several days, techniques in the field of event identification might be used to map a certain piece of news to this development. Then, it could also be analysed how this certain piece of news ties in with several concurrent, and related news articles, and how these might influence the frequency increase of on and off-topic comments respectively. This might grant more insight into key events within discussion on a certain topic. However, frequency alone might be too limited with regard to the fact that it allows only research that relates to increase of discussion. Delving deeper into the content of the classified documents might give an indication on how certain events alter the content of a given topic.

An example of this is Sentiment Analysis. The research that has frequently been conducted on social media to give a shallow form of public opinion by using classification of sentiment (positive, neutral, and negative for example) could very well be integrated in a more extensive theoretical framework. In line with the subject of this thesis, as well as the extensions that have been discussed up until now, this might take form as follows: as a first effort, it could be analysed how the topic of privacy, being the distribution over words, is built up in terms of sentiment. Is the topic a priori predominantly negatively or positively associated? How do sources differ in this, and is there a change in these weights overtime? Moreover, it could be assessed which events lead to positive and negative fluctuations in the sentiment weights over time. In this sense, it could be hypothesized more concretely that the actions of the NSA have increased the sentiment on the topic of privacy negatively. From there, it might even be possible to analyse which other topics are simultaneously influenced negatively by these actions. It might be the case that, for example, certain companies or services linked to PRISM might display an increase of negative sentiment after the program's workings having leaked.

"So this is the nullspace of A, or $NS(A)\dots$ the NSA is everywhere."

– Differential equations professor

A LINEAR ALGEBRA OPERATIONS

A.1 MATRIX TERMINOLOGY

Say we have a matrix A (depicted in Example 18), a *transpose* matrix is then that of A its rows converted into columns, superscripted by ^{T}. The *identity* matrix I is an $n \cdot n$ matrix, with a diagonal of all 1's, that if multiplied by $m \cdot n$ matrix A with gives $I_m A = A$ and $AI_n = A$. Therefore, it is as if multiplying a given matrix by one. This special feature of the identity matrix can be used to find an orthogonal matrix; if $AA^T = A^T A = I$, it is orthogonal. Lastly, to understand the process of finding Eigenvectors and Eigenvalues, one needs to know what a determinant is.

This chapter intends to provide a brief overview of the terms from linear algebra used to explain the models in this thesis.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad A^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{18}$$

The determinant of a matrix A, denoted by $|A|$ or $\det(A)$, can be used to reduce it to a single number. For a $2 \cdot 2$ matrix, this can be done by:

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \tag{19}$$

This becomes increasingly harder with larger matrices, which then need to be split into $2 \cdot 2$ matrices in order to compute the determiner:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = (a)\begin{vmatrix} e & f \\ h & i \end{vmatrix} - (b)\begin{vmatrix} d & f \\ g & i \end{vmatrix} + (c)\begin{vmatrix} d & e \\ g & h \end{vmatrix} \tag{20}$$

A.2 EIGENVECTORS AND EIGENVALUES

The splitted matrix can be used to multiply the vector $\vec{x}$; we put in and out comes vector $A\vec{x}$. Recall that we looked at how vectors are composed in section 3.1.3. It can be any list of numbers, or words, that form the attributes of a given instance, in for example a VSM, where these attributes determine the weight and therefore the direction of a given vector. So, the vector $\vec{x}$ points in a certain direction in a high dimensional space and A means to change this direction by adjusting the attributes. Now, there are certain vectors where $A\vec{x}$ comes out parallel (in the same direction) to $\vec{x}$, which is then called an Eigenvector. Formally: $A\vec{x} = \lambda\vec{x}$, where λ is the Eigenvalue with $A\vec{x}$ being the Eigenvector. So given:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \rightarrow \quad A\vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \lambda = 1 \tag{21}$$

The same also holds for:

$$\vec{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \rightarrow \quad A\vec{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \lambda = -1 \tag{22}$$

This implies that $n \cdot n$ matrices also have n Eigenvectors, and, that $\sum \forall\lambda \in A = a_{11}, a_{22}, \ldots, a_{nn}$. This works the other way around as well; to find the Eigenvalues, we note:

$$\begin{aligned} Ax &= \lambda x \\ (A - \lambda I)x &= 0 \\ \det(A - \lambda I) &= 0 \end{aligned} \tag{23}$$

So for example:

$$\begin{aligned} A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad \det(A - \lambda I) &= \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} \\ &= (3-\lambda)^2 - 1 \\ &= \lambda^2 - 6\lambda + 8 \\ &= \frac{(\lambda - 4)}{(\lambda - 2)} \end{aligned} \tag{24}$$

Now that it is known that $\lambda_1 = 4$ and $\lambda_2 = 2$, which are our Eigenvalues, it is possible to find the Eigenvectors by subtracting λ_1 and λ_2 from A individually.

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \quad A - 4I = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{25}$$

Note that the subtraction is done for the $a_{n,n}$ elements in the diagonal Matrix 26, which is the *trace* denoted by $\mathrm{tr}(A)$.

$$A_{n,n} = \begin{bmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{bmatrix} \tag{26}$$

A.3 NULL SPACE

The 1's in the Null Space form the vector that matches x_1 in Example 23, which can be found through $N(A) = \{\vec{x} \in \mathbb{R} | A\vec{x} = \vec{0}\}$. So, its a case of finding all $\vec{x}$'s that satisfy $A\vec{x} = \vec{0}$. Going into detail on finding the Null Space (or Kernel) of a matrix requires knowledge about reduced row-echelon form (denoted by $\mathrm{rref}(A)$).

A.3.1 *Reduced Row-Echelon Form*

A given matrix A is in $\mathrm{rref}(A)$ if all rows fully populated with zeros are at the bottom and there is a diagonal of first non-zero entries (pivots), therefore it being in row-echelon form, which is *also* the only non-zero in its column. An example taken from Sadun (2008):

$$\begin{gathered} A = \begin{bmatrix} 2 & -2 & 4 & -3 \\ 2 & 1 & 10 & 7 \\ -4 & 4 & -8 & 4 \\ 4 & -1 & 14 & 6 \end{bmatrix} \rightarrow A_r = \begin{bmatrix} 2 & -2 & 4 & -2 \\ 0 & 3 & 6 & 9 \\ 0 & 0 & 0 & 0 \\ 0 & 3 & 6 & 10 \end{bmatrix} \\ A_{\mathrm{ref}} = \begin{bmatrix} 2 & -2 & 4 & -2 \\ 0 & 3 & 6 & 9 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow A_{\mathrm{rref}} = \begin{bmatrix} 1 & 0 & 4 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{gathered} \tag{27}$$

The matrix A in Example 27 can be converted to row-echelon form by using Gaussian elimination, or row reduction. These are the steps to arrive at A_r:

(1) $A_2 = A_2 - A_1$

(2) $A_3 = A_3 + A_1 \cdot 2$

(3) $A_4 = A_4 - A_1 \cdot 2$

The matrix now looks like A_r.

(4) $A_4 = A_4 - A_2$

(5) Swap A_3 and A_4.

We now have matrix A in row-echelon form A_{ref}. To get the reduced row-echelon form, we continue:

(6) $A_1 = \frac{A_1}{2}$

(7) $A_2 = \frac{A_2}{3}$

(8) $A_1 = A_1 + A_3$

(9) $A_2 = A_2 - A_3 \cdot 3$

(10) $A_1 = A_1 + A_2$

This yields the matrix A_{rref}; A in reduced row-echelon form. On a side note, we can now determine the rank of either A, A_{ref} or A_{rref} by the number of their pivots.

A.3.2 *Calculate Null Space with RREF*

Now that we know how the reduced row-echelon form works, we can find the basis of the Null Space $Ax = 0$. We know A and have calculated A_{rref}, then, given the equations:

$$\begin{aligned} a_1 1x_1 + a_1 2x_2 + \ldots + a_1 n x_n &= 0 \\ a_2 1x_1 + a_2 2x_2 + \ldots + a_2 n x_n &= 0 \\ &\vdots \\ a_m 1x_1 + a_m 2x_2 + \ldots + a_m n x_n &= 0 \end{aligned} \tag{28}$$

Now we need to correctly read our equations off A_{rref}, namely:

$$\begin{aligned} x_1 + 4x_3 &= 0 \\ x_2 + 2x_3 &= 0 \\ x_4 &= 0 \\ 0 &= 0 \end{aligned} \tag{29}$$

What we did is check the *pivot variables* for the rows x_1, x_2 and x_4, as these are *constrained variables* which have to be in our list of non-trivial equations shown in Example 29. Because x_3 is a *free variable*, we can plug it into our trivial equations and remove the $0 = 0$ bit, then we get:

$$\begin{aligned} x_1 &= -4x_3 \\ x_2 &= -2x_3 \\ x_3 &= x_3 \\ x_4 &= 0 \end{aligned} \tag{30}$$

Now, this set is our solution to $Ax = 0$, which is then all multiples c of the set, giving:

$$x = c \cdot \begin{bmatrix} -4 \\ -2 \\ 1 \\ 0 \end{bmatrix}^T \tag{31}$$

An additional example can be found in Sadun (2008). This should provide enough information regarding the elements that underlie Singular Value Decomposition.

A.4 SINGLULAR VALUE DECOMPOSITION

Assume we have a term · document matrix; combining a frequency list such as that from previous Table **??** as rows, and the documents a columns. Note that in contrast to the example in Table **??**, there can now be zero values where a words does not occur in a certain document, and that the matrix is not ranked on frequency. We then take:

$$A = \begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix} \tag{32}$$

Which means that AA^T will be:

$$AA^T = \begin{bmatrix} 2 & 0 & 8 & 6 & 0 \\ 1 & 6 & 0 & 1 & 7 \\ 5 & 0 & 7 & 4 & 0 \\ 7 & 0 & 8 & 5 & 0 \\ 0 & 10 & 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} 2 & 1 & 5 & 7 & 0 \\ 0 & 6 & 0 & 0 & 10 \\ 8 & 0 & 7 & 8 & 0 \\ 6 & 1 & 4 & 5 & 0 \\ 0 & 7 & 0 & 0 & 7 \end{bmatrix} = \begin{bmatrix} 104 & 8 & 90 & 108 & 0 \\ 8 & 87 & 9 & 12 & 109 \\ 90 & 9 & 90 & 111 & 0 \\ 108 & 12 & 111 & 138 & 0 \\ 0 & 109 & 0 & 0 & 149 \end{bmatrix}$$

(33)

With the orthonormal eigenvectors of AA^T, we can determine the column vectors for U: $\lambda = 321.07, \lambda = 230.17, \lambda = 12.70, \lambda = 3.49, \lambda = 0.12$. And then compute and order the matrix:

$$U = \begin{bmatrix} -0.54 & 0.07 & 0.82 & -0.11 & 0.12 \\ -0.10 & -0.59 & -0.11 & -0.79 & -0.06 \\ -0.53 & 0.06 & -0.21 & 0.12 & -0.81 \\ -0.65 & 0.07 & -0.51 & 0.06 & 0.56 \\ -0.06 & -0.80 & 0.09 & 0.59 & 0.04 \end{bmatrix} \tag{34}$$

Looking at Example 34 already reveals some information regarding the relations words share; the first column vector is all negative, which implies that all words co-occur in each document. Another observation which can be made is that $w_{2,2}$ co-occurs with $w_{2,5}$, $w_{3,2}$ co-occurs with $w_{3,4}$ and $w_{3,5}$ and so on. For calculating V^T, we repeat this process for $A^T A$ and apply the Gram-Schmidt orthonormalization process Baker (2005), which converts a set of vectors to their orthonormal form by normalizing a given vector and iteratively rewriting the remaining vectors by multiplying themselves with the normalized vectors:

$$A^T A = \begin{bmatrix} 79 & 6 & 107 & 68 & 7 \\ 6 & 136 & 0 & 6 & 112 \\ 107 & 0 & 177 & 116 & 0 \\ 68 & 6 & 116 & 78 & 7 \\ 7 & 112 & 0 & 7 & 98 \end{bmatrix} \tag{35}$$

$$V^T = \begin{bmatrix} -0.46 & 0.02 & -0.87 & -0.00 & 0.17 \\ -0.07 & -0.76 & 0.06 & 0.60 & 0.23 \\ -0.74 & 0.10 & 0.28 & 0.22 & 0.56 \\ -0.48 & 0.03 & 0.40 & -0.33 & 0.70 \\ -0.07 & -0.64 & -0.04 & -0.69 & -0.32 \end{bmatrix} \tag{36}$$

If we again take the eigenvalues from this matrix we get the linear independent components, indicating the amount of variance per dimension. If we only take three of these components for S, it will yield:

$$S = \begin{bmatrix} 17.92 & 0 & 0 \\ 0 & 15.17 & 0 \\ 0 & 0 & 3.56 \end{bmatrix} \tag{37}$$

This can be done in a similar fashion for U and V^T, therefore yielding the full $A = U\Sigma V^T$:

$$
\begin{aligned}
S &= \\
&\begin{bmatrix} -0.54 & 0.07 & 0.82 \\ -0.10 & -0.59 & -0.11 \\ -0.53 & 0.06 & -0.21 \\ -0.65 & 0.07 & -0.51 \\ -0.06 & -0.80 & 0.09 \end{bmatrix}
\begin{bmatrix} 17.92 & 0 & 0 \\ 0 & 15.17 & 0 \\ 0 & 0 & 3.56 \end{bmatrix}
\begin{bmatrix} -0.46 & 0.02 & -0.87 & -0.00 & 0.17 \\ -0.07 & -0.76 & 0.06 & 0.60 & 0.23 \\ -0.74 & 0.10 & 0.28 & 0.22 & 0.56 \end{bmatrix} \\
&= \begin{bmatrix} 2.29 & -0.66 & 1.25 & -3.09 \\ 1.17 & 6.76 & -5.50 & -2.13 \\ 4.86 & -0.96 & 0.38 & -0.97 \\ 6.62 & -1.23 & 0.24 & -0.71 \\ 1.14 & 9.19 & -7.19 & -3.13 \end{bmatrix}
\end{aligned}
\tag{38}
$$

B RETRIEVAL

This Appendix provides a very shallow introduction on how the documents for this thesis were retrieved and the limitations of this process.

B.1 SCRAPING

Scrapers are programs that visit websites and typically 'scrape' the desired content from the page they are visiting. Through this, they simulate a human visiting the actual websites by using simple HTTP requests or even a full browser. Scrapers can function by recursively iterating through all links from a specific URL, or more directly visit a range of content, by article number for example. As queries issued by these scrapers generally only take milliseconds of time to retrieve a website its page, they can be used to fill complete datasets in only days time. At first glance, these programs might be considered an ideal way to extract information; however, due to the fact that these scrapers are able to put a tremendous load on servers hosting websites, server-sided counter measures to prevent overloading are common.

B.2 PARSING

After a scraper has fetched the content from a website it can be useful to extract only the information required, which in case of language processing usually concerns raw text. Websites, however, are written in a markup language called `HTML`. It supplies tags by which elements, such as a block of text or an image, can be produced, as well as identified with. Although the tags pollute the document with tags, a parser can effectively make use of these to only extract the required information. Take for example Figure 26. An HTML parser would allow for extracting the information in from a certain tag. Say for example that we want to extract the date from an article from which the above example is a small excerpt. By telling the parser to "deliver text from the `div` that has `"dateplace-data"` as `class`", it then delivers the date. In combination with a scraper automatically delivering web documents, these can be delivered to a database in a structured fashion, optimizing analysis possibilities.

```
<div class="header" >
  <div class="dateplace" >
    <div class="dateplace-data" >
      2 december 2013 18:33
      2 december 2013 18:36 <br />
    </div>
  </div>
</div>
```

Figure 26: Example HTML structure.

B.3 LIMITATIONS

One of the simplest considerations of scraping lies in how the articles can be visited. If a news site does not number-label its article entries, it becomes almost impossible to retrieve them if they do not have an accessible archive to extract the titles from. An ideal URL composition would be the following:

```
http://www.website.something/<article number>
```

This would allow the articles to be easily visited with a range of possible numbers alone. Alas, not every website is constructed with optimization for retrieval in mind. This becomes even more evident when looking under the hood. Websites with a consistent structure, in this example those where the date can always be found in a `div` where `class="dateplace-data"`, make retrieval in terms of parsing much less complicated. Another stumbling block that is yielded by this general lack of structure is retrieval time. If a program that automatically extracts these articles is imagined, for this example not regarding any limitations the website imposes on how quickly this program can take actions, an optimized article page would have a flat structure. By this is meant one page per article with all the content, such as title, body, and comments, that should be extracted within this page. Any step added to this rapidly increases computation and therefore retrieval time. If we do regard website-sided limitations on the program used for retrieval, this becomes even trickier. Finally, as mentioned, scrapers can place an abnormal load on the servers hosting these websites, resulting in sites taking counter measures to recognize and blacklist these systems. If obfuscating a scraper for a website affects the computation time severely, the retrieval time again increases.

B.4 DATA STRUCTURE

To be able to use each document instance extracted by the scraper it has to be represented in a format that allows for a clear structure and database integration for faster and simple comparison between instances. The JSON format can be used to do exactly this; it represents each document as an object with attributes, which in turn can also have attributes (Figure 27). For example, a document has an id and tags, which have a single value, and comments, which in turn have their own id, author and text. These objects can be stored in a NoSQL database, which derives the structure of the database from the classes, rather than the structure having to be preformed as in traditional database structures.

```json
{
    "id": "extraction number",
    "source": "t.net / nu.nl",
    "nr": "article number",
    "date": "date",
    "year": "year",
    "time": "time",
    "subjects": "subject 1, subject 2, etc",
    "content": {
        "title": "title",
        "intro": "introduction",
        "text": "text"
    },
    "comments": [
        {   "comment_id": "id",
            "comment_user": "user name",
            "comment_date": "date",
            "comment_year": "year",
            "comment_time": "time",
            "comment_text": "text",
            "comment_vote": "vote nr"
        },
        {   "comment_id": "id",
            "comment_user": "user name",
            "comment_date": "date",
            "comment_year": "year",
            "comment_time": "time",
            "comment_text": "text",
            "comment_vote": "vote nr",
        }
    ]
}
```

Figure 27: A document instance structured in JSON format.

BIBLIOGRAPHY

Adermon, A., & Liang, C.-Y. (2014, May). Piracy and Music Sales: The Effects of an Anti-Piracy Law. *Journal of Economic Behavior & Organization*. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0167268114001395 doi: 10.1016/j.jebo.2014.04.026

Appelbaum, J. (2013). *To Protect And Infect - Part 2.* CCCen. Retrieved from www.youtube.com/watch?v=vILAlhwUgIU

Asuncion, A. (2010). Approximate mean field for Dirichlet-based models. *ICML Workshop on Topic Models*(i). Retrieved from http://mmm.csd.uwo.ca/faculty/yuri/Papers/cvpr10.pdf

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. (2009). On smoothing and inference for topic models. *. . . Twenty-Fifth Conference on . . .* (Ml). Retrieved from http://dl.acm.org/citation.cfm?id=1795118

Baker, K. (2005). Singular value decomposition tutorial. *The Ohio State University*, *2005*, 1–24. Retrieved from http://lsa-svd-application-for-analysis.googlecode.com/svn-history/r120/trunk/LSA/Other/LsaToRead/SVDTut.pdf

Barlow, J. P. (1996). A Declaration of the Independence of Cyberspace, February 8, 1996. Retrieved from http://homes.eff.org/barlow/Declaration-Final.html

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. *ICWSM*, 438–441. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2745/3207

Bellman, S., Johnson, E. J., Kobrin, S. J., & Lohse, G. L. (2004). International differences in information privacy concerns: A global survey of consumers. *The Information Society*, *20*(5), 313–324. Retrieved from http://www8.gsb.columbia.edu/sites/decisionsciences/files/files/1172.pdf

Bennett, C. J. (1995). The political economy of privacy: a review of the literature. *center for social and legal research, DOE genome project (Final draft), University of Victoria, Department of Political Science, Victoria.*

Berendt, B., Günther, O., & Spiekermann, S. (2005, April). Privacy in e-commerce. *Communications of the ACM*, *48*(4), 101–106. Retrieved from http://portal.acm.org/citation.cfm?doid=1053291.1053295http://www.wiwi.uni-siegen.de/itsec/publikationen/19.pdf doi: 10.1145/1053291.1053295

Berger, A., Pietra, V., & Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational linguis-*

tics(1992). Retrieved from http://dl.acm.org/citation.cfm?id=234289

Blei, D. (2010, April). Probabilistic Topic Models. *Communications of the ACM*, 77–84. Retrieved from http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf doi: 0.1145/2133806.2133826

Blei, D., & McAuliffe, J. (2007). Supervised Topic Models. *NIPS*, 1–8. Retrieved from https://papers.nips.cc/paper/3328-supervised-topic-models.pdf

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The journal of machine learning research*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944937

Blei, D. M. (2009). *Topic Models.* Retrieved 22-04-2014, from http://videolectures.net/mlss09uk_blei_tm/

Buchanan, T., & Paine, C. (2007). Development of measures of online privacy concern and protection for use on the Internet. *Journal of the American Society for Information Science and Technology*, *58*(2), 157–165. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.20459/full doi: 10.1002/asi

Burtion, M. (2013). *The Joy of Topic Modeling.* Retrieved 1-5-2014, from http://mcburton.net/blog/joy-of-tm/

Chellappa, R., & Sin, R. (2005, April). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management*, *6*(2-3), 181–202. Retrieved from http://link.springer.com/10.1007/s10799-005-5879-yhttp://link.springer.com/article/10.1007/s10799-005-5879-y doi: 10.1007/s10799-005-5879-y

Consumers-Union. (2008). "Consumer Reports Poll: Americans Extremely Concerned About Internet Privacy". Retrieved from http://www.consumersunion.org/pub/core_telecom_and_utilities/006189.html

Deerwester, S., Dumais, S., & Landauer, T. (1990). Indexing by latent semantic analysis. *JASIS*. Retrieved from http://www.cob.unt.edu/itds/faculty/evangelopoulos/dsci5910/LSA_Deerwester1990.pdf

Der Spiegel. (2013). *NSA Spying Scandal.* Retrieved from http://www.spiegel.de/international/topic/nsa_spying_scandal/

Drennan, J., Sullivan, G., & Previte, J. (2006). Privacy, risk perception, and expert online behavior: an exploratory study of household end users. *Journal of Organizational and End User Computing (JOEUC)*, *18*(1), 1–22. Retrieved from http://www.igi-global.com/article/privacy-risk-perception-expert-online/3806

Dutton, W., Genarro, C. D., & Millwood, A. (2005). *The internet in Britain* (No. May). Retrieved from http://live.online.se/wip/publishedarchive/oxis2005_report.pdf

Etzioni, A. (2014). *NSA: National Security vs. Individual Rights* (Vol. 00) (No. 0). Taylor & Francis. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02684527.2013.867221 doi: 10.1080/02684527.2013.867221

Fain, D., & Pedersen, J. (2006). Sponsored search: A brief history. *Bulletin of the American Society for . . .* (818). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/bult.1720320206/full

FCC. (2010). *Free Internet Order* (Tech. Rep.).

FCC. (2014). *Notice of Proposed Rulemaking* (Tech. Rep.). Retrieved from http://transition.fcc.gov/Daily_Releases/Daily_Business/2014/db0515/FCC-14-61A1.pdf

Forrester Research. (2005). *No Title.* Retrieved from http://www.forrester.com/Research/Document/Excerpt/0,7211,38299,00.htmlGalegher,

Fuchs, C. (2011). Towards an alternative concept of privacy. *Journal of Information, Communication and Ethics in Society*, *9*(4), 220–237. Retrieved from http://www.emeraldinsight.com/10.1108/14779961111191039 doi: 10.1108/14779961111191039

Giddens, A. (2013). *The constitution of society: Outline of the theory of structuration*. John Wiley & Sons.

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. *Advances in Knowledge Discovery and Data . . .*, 22–30. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24775-3_5

Griffiths, T. L., & Steyvers, M. (2004, April). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101 Suppl*, 5228–35. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=387300&tool=pmcentrez&rendertype=abstract doi: 10.1073/pnas.0307752101

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks. *. . . 2005 ACM workshop on Privacy in the electronic . . .* . Retrieved from http://dl.acm.org/citation.cfm?id=1102214

Hann, I., Hui, K., Lee, S., & Png, I. (2002). Online Information Privacy: Measuring the Cost-Benefit Trade-Off. *ICIS*, 1–10. Retrieved from http://www.comp.nus.edu.sg/~ipng/research/privacy_icis.pdf

Harris Inc. (2004). New National Survey on Consumer Privacy Attitudes. In *Privacy & american bussiness landmark conference.*

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 50–57. Retrieved from http://portal.acm.org/citation.cfm?doid=312624.312649 doi: 10.1145/312624.312649

Instagram. (2013). *Terms of Use.* Retrieved from http://instagram.com/about/legal/terms/

Jensen, C., & Potts, C. (2004). Privacy policies as decision-making tools: an evaluation of online privacy notices. *Proceedings of the SIGCHI conference on Human . . .* , *6*(1), 471–478. Retrieved from http://dl.acm.org/citation.cfm?id=985752

Joinson, A., & Reips, U. (2010). Privacy, trust, and self-disclosure online. *Human-Computer Interaction*, 1–46. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/07370020903586662

Jupiter Research. (2002). Security and privacy data. *Security and privacy data. Presentation to the Federal Trade Commission Consumer Information Security*. Retrieved from http://www.ftc.gov/bcp/workshops/security/0205201leathern.pdf

Kaur, A., & Gupta, V. (2013, November). A Survey on Sentiment Analysis and Opinion Mining Techniques. *Journal of Emerging Technologies in Web Intelligence*, *5*(4), 367–371. Retrieved from http://ojs.academypublisher.com/index.php/jetwi/article/view/11660 doi: 10.4304/jetwi.5.4.367-371

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *1*(2), 211–240. Retrieved from http://psycnet.apa.org/journals/rev/104/2/211/

Lanier, J. (2013). *Who owns the future?* Simon and Schuster.

Leiner, B., Cerf, V., & Clark, D. (2009). A brief history of the Internet. *ACM SIGCOMM . . .* , *39*(5), 22–31. Retrieved from http://dl.acm.org/citation.cfm?id=1629613

Liebowitz, S. (2012). Policing pirates in the networked age. (438). Retrieved from http://www.cato.org/publications/policy-analysis/policing-pirates-networked-age

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge university press. Retrieved from http://www.langtoninfo.co.uk/web_content/9780521865715_frontmatter.pdf

McGowan, C. (2014). The Relevance of Relevance: Section 215 of the USA PATRIOT Act and the NSA Metadata Collection Program. *Fordham Law Review*, *82*(5). Retrieved from http://ir.lawnet.fordham.edu/flr/vol82/iss5/15/

Metzger, M. J. (2006). Effects of Site, Vendor, and Consumer Characteristics on Web Site Trust and Disclosure. *Communication Research*, *33*(3), 155–179.

Milne, G. R., & Culnan, M. J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing*, *18*(3), 15–29.

O'Connor, B., & Balasubramanyan, R. (2010). From tweets to polls:

Linking text sentiment to public opinion time series. *ICWSM*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPDFInterstitial/1536/1842

OECD. (1999). *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.* Retrieved 5-6-2013, from bu

Orwell, G. (1949). *1984*. Editions Underbahn Ltd.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. , 1–17. Retrieved from http://ilpubs.stanford.edu:8090/422

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceeding of the 17th international conference on World Wide Web - WWW '08*, 91. Retrieved from http://portal.acm.org/citation.cfm?doid=1367497.1367510 doi: 10.1145/1367497.1367510

Quercia, D., Askham, H., & Crowcroft, J. (2012). TweetLDA: supervised topic classification and link prediction in Twitter. *Proceedings of the 3rd Annual ACM ...*, 1–4. Retrieved from http://dl.acm.org/citation.cfm?id=2380750

Ramage, D., & Hall, D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 ...* (August), 248–256. Retrieved from http://dl.acm.org/citation.cfm?id=1699543

Roba, R. (2009). The Legal Protection Of The Secrecy Of Correspondence. *Curentul Juridic, The Juridical Current, Le Courant ...* (1272008). Retrieved from http://www.upm.ro/facultati_departamente/ea/RePEc/curentul_juridic/rcj09/recjurid091_10F.pdf

Rosen, J. (2011). *The unwanted gaze: The destruction of privacy in America*. Random House LLC.

Rustad, M. (2011). The Path of Internet Law: An Annotated Guide To Legal Landmarks. *Duke L. & Tech. Rev.*. Retrieved from http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1226&context=dltr

Sadun, L. (2008). *Matrix Operations.* Retrieved 29-03-2014, from https://www.ma.utexas.edu/users/sadun/S08/427K/matrix.pdf

Salakhutdinov, R., Tenenbaum, J., & Torralba, A. (2011). Learning to Learn with Compound HD Models. *NIPS*, 1–9. Retrieved from http://books.nips.cc/papers/files/nips24/NIPS2011_1163.spotlight.pdf

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11). Retrieved from http://dl.acm.org/citation.cfm?id=361220

Sato, I., & Nakagawa, H. (2012). Rethinking collapsed variational Bayes inference for LDA. *arXiv preprint arXiv:1206.6435*. Retrieved from http://arxiv.org/abs/1206.6435

Sheehan, K. B. (2002). Toward a Typology of Internet Users and Online Privacy Concerns. *The Information Society*, *18*(1), 21–32. Retrieved from http://dx.doi.org/10.1080/01972240252818207 doi: 10.1080/01972240252818207

Smith, H., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS quarterly*, *35*(4), 989–1015. Retrieved from http://dl.acm.org/citation.cfm?id=2208950

Solove, D. J. (2004). *The digital person: Technology and privacy in the information age* (Vol. 1). NYU Press.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, 477. Retrieved from http://www.jstor.org/stable/40041279

Solove, D. J. (2011). *Nothing to hide: The false tradeoff between privacy and security*. Yale University Press.

Sorower, M. (2010). A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 1–25. Retrieved from http://people.oregonstate.edu/~sorowerm/pdf/Qual-Multilabel-Shahed-CompleteVersion.pdf

Spiekermann, S., Grossklags, J., & Berendt, B. (2001). E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. *Proceedings of the 3rd ACM conference on Electronic Commerce*. Retrieved from http://dl.acm.org/citation.cfm?id=501163

Srivastava, A., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. CRC Press.

Stephenson, N. (2011). *README*. Atlantic Books Ltd.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*.

The Guardian. (2013). *The NSA Files.* Retrieved 5-5-2014, from http://www.theguardian.com/world/the-nsa-files

Tokunaga, T., & Makoto, I. (1994). Text categorization based on weighted inverse document frequency. *Special Interest Groups and Information Process . . .* . Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7015

Tweakers.net. (2010a). *Gevolgen verkiezingen voor privacy hangen af van coalitie.* Retrieved from https://tweakers.net/nieuws/67872/gevolgen-verkiezingen-voor-privacy-hangen-af-van-coalitie.html

Tweakers.net. (2010b). *VVD wil kleinere databanken en betere bescherming privacy.* Retrieved from https://tweakers.net/nieuws/67135/vvd-wil-kleinere-databanken-en-betere-bescherming-privacy.html

Tweakers.net. (2012a). *Teeven: groot deel gemelde datalekken kan niet onderzocht worden.* Retrieved from https://tweakers.net/nieuws/80982/teeven-groot-deel-gemelde-datalekken-kan

-niet-onderzocht-worden.html

Tweakers.net. (2012b). *Teeven wil thuiskopieheffing mogelijk gaan uitbreiden.* Retrieved from https://tweakers.net/nieuws/82937/teeven-wil-thuiskopieheffing-mogelijk-gaan-uitbreiden.html

Tweakers.net. (2013a). *CBP: overheid laks bij bescherming burgergegevens.* Retrieved from https://tweakers.net/nieuws/88589/cbp-overheid-laks-bij-bescherming-burgergegevens.html

Tweakers.net. (2013b). *Datahonger: de Nederlandse geheime dienst wil alles weten.* Retrieved 9-5-2014, from https://tweakers.net/reviews/3308/1/datahonger-de-nederlandse-geheime-dienst-wil-alles-weten-inleiding.html

Tweakers.net. (2013c). *Nieuwe wetgeving: Nederland bouwt aan zijn eigen NSA.* Retrieved 10-5-2014, from https://tweakers.net/reviews/3337/nieuwe-wetgeving-nederland-bouwt-aan-zijn-eigen-nsa.html

Tweakers.net. (2013d). *Xs4all en Ziggo: aanpakken filesharers werkt beter dan Pirate Bay-blokkade.* Retrieved 11-5-2014, from https://tweakers.net/nieuws/91387/xs4all-en-ziggo-aanpakken-filesharers-werkt-beter-dan-pirate-bay-blokkade.html

Tweakers.net. (2014a). *Kabinet: zet telefoon uit om wifi-tracking tegen te gaan.* Retrieved from https://tweakers.net/nieuws/94273/kabinet-zet-telefoon-uit-om-wifi-tracking-tegen-te-gaan.html

Tweakers.net. (2014b). *Minister Plasterk overleeft motie van wantrouwen over verwarring metadata.* Retrieved 10-5-2014, from https://tweakers.net/nieuws/94262/minister-plasterk-overleeft-motie-van-wantrouwen-over-verwarring-metadata.html

Tweakers.net. (2014c). *Regering: optie om kijkgedrag te monitoren hoeft standaard niet uit bij smart-tv.* Retrieved from https://tweakers.net/nieuws/94476/regering-optie-om-kijkgedrag-te-monitoren-hoeft-standaard-niet-uit-bij-smart-tv.html

Tweede Kamer der Staten-Generaal. (2013). *Kabinetsbrede reactie onthullingen Snowden* (Tech. Rep.).

Uerpmann-Wittzack, R. (2010). Principles of international internet law. *German LJ*, *367*. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/germlajo11&section=86

Vavliakis, K. N., Symeonidis, A. L., & Mitkas, P. a. (2013, November). Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering*, *88*, 1–24. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169023X13000827 doi: 10.1016/j.datak.2013.08.006

Vlemmix, P. (2012). *Panopticon.* Retrieved from http://www.panopticondefilm.nl

Wang, C.-J., Wang, P.-P., & Zhu, J. J. H. (2013, September). Discussing occupy wall street on Twitter: longitudinal network analysis of equality, emotion, and stability of public discussion. *Cyberpsychology, behavior and social networking*, *16*(9), 679–85. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23656222 doi: 10.1089/cyber.2012.0409

Westin, A. F., & Blom-Cooper, L. (1970). *Privacy and freedom* (Vol. 67). Atheneum New York.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

LIST OF FIGURES

LIST OF TABLES