Universität Potsdam
Faculty of Economics and Social Sciences
Chair of Business Information Systems,
esp. Social Media and Society

Toxicity and Disorder: Investigating Users' Online Behaviour Through the Lens of the Broken Windows Theory

Master Thesis submitted by

Anthony Hamilton Fernando

in fulfilment of the requirements for the degree of

Master of Science

in the program

Wirtschaftsinformatik und Digitale Transformation

1. Supervisor:	**Dr. Annika Baumann**
2. Supervisor:	**Prof. Dr. Hanna Krasnova**
Matriculation Number:	816086
Date:	May 22, 2024

Acknowledgement

This thesis would not have been possible without the support of many people around me. First of all, I would like to thank Dr. Annika Baumann, who was a great supervisor, because she gave me the right ideas to get the best out of this thesis. I am very grateful for her support in general, but especially for her rapid responses and her open discussions throughout the whole process. Moreover, I would like to thank Prof. Dr. Hanna Krasnova for being my second reviewer. I would also like to express my gratitude to Moritz Wilksch, who provided me with the LaTeX template for this thesis. Additionally, I would like to thank Dr. Media Khorasani, my working colleague from the BAM, for the support and for cheering me up throughout the whole process. Last but not least, my thanks go to my parents and friends, especially Janis, Melissa, Valentin and Miru for supporting and motivating me the whole period.

Abstract

The usage of online environments such as social media has grown steadily, while at the same time, the toxic commenting behaviour in this environment has dramatically increased. Many factors related to toxic behaviour, such as anonymity, have been well studied, whereas factors like the effect of a disordered website design have been less studied. In offline environments, however, disorder and antisocial behaviour have been researched intensively, which is also known as the Broken Windows Theory (BWT). This thesis aims to transfer the knowledge of the BWT from offline to online environments to determine if perceived disorder in an online environment increases toxic commenting behaviour. To accomplish this aim, the social media platform Reddit was selected as the online environment to be analysed. It had a major design change from a disordered to a less disordered website design, where a "before" and "after" analysis of two subreddits (r/relationships and r/relationship_advice) with the same topic was completed. In addition to the Natural Language Processing methods, four approaches were employed to study the aim: Descriptive, content-based, user-based and submission-based. Based on those analyses, it could be seen that "after" the redesign, the toxic commenting behaviour in only one subreddit (r/relationships) has decreased. However, the highest impact on the toxic behaviour was found in the moderation rules of each subreddit. This implies that the BWT can not be fully applied in the online environment for this subreddit pair.

Keywords: Broken Windows Theory, Toxicity, Perceived Disorder, Online Environment, Reddit, Commenting Behaviour.

Zusammenfassung

Die Nutzung von Online-Umgebungen wie den sozialen Medien hat stetig zugenommen, während gleichzeitig das toxische Kommentarverhalten in dieser Umgebung dramatisch zugenommen hat. Viele Faktoren, die mit toxischem Verhalten in Zusammenhang stehen, wie z.B. Anonymität, sind gut erforscht, während Faktoren wie die Auswirkungen einer unordentlichen Gestaltung einer Website weniger untersucht wurden. In der Offline-Umgebung hingegen sind Unordnung und antisoziales Verhalten intensiv erforscht worden, was auch als Broken Windows Theory (BWT) bekannt ist. Ziel dieser Arbeit ist es, die Erkenntnisse der BWT von der Offline- auf die Online-Umgebung zu übertragen, um festzustellen, ob die wahrgenommene Unordnung in einer Online-Umgebung das toxische Kommentarverhalten erhöht. Um dieses Ziel zu erreichen, wurde die Social-Media-Plattform Reddit als zu untersuchende Online-Umgebung ausgewählt. Das Design der Plattform wurde von einem ungeordneten zu einem weniger ungeordneten Website-Design geändert, und es wurde eine "Vorher"-"Nachher"-Analyse von zwei Subreddits (r/relationships und r/relationship_advice) mit demselben Thema durchgeführt. Zusätzlich zu den Methoden der natürlichen Sprachverarbeitung wurden vier Ansätze zur Untersuchung des Ziels verwendet: Descriptive, Content-basiert, User-basiert und Submission-basiert. Anhand dieser Analysen konnte festgestellt werden, dass "nach" der Umgestaltung das toxische Kommentarverhalten in nur einem Subreddit (r/relationships) zurückgegangen ist. Der größte Einfluss auf das toxische Verhalten wurde jedoch bei den Moderationsregeln der einzelnen Subreddits festgestellt. Dies deutet darauf hin, dass die BWT in der Online-Umgebung für dieses Subreddit-Paar nicht vollständig angewendet werden kann.

Schlüsselwörter: Broken Windows Theory, Toxizität, Wahrgenommene Unordnung, Online Umgebung, Reddit, Kommentarverhalten.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Problem

In recent years, new environments were developed that are called "online environments". Online environments, mainly social media platforms, are one of the most innovative developments of the 21st century (Sheth et al., 2022). This innovation allows individuals to communicate, interact and share their opinions around the world in real-time with other individuals (Fan et al., 2021; S. Kumar et al., 2017). It is seen as breaking geographical boundaries (Sheth et al., 2022). This environment makes individuals feel like building a society that includes a supportive network by connecting and communicating through comments with other users worldwide (Abualigah et al., 2020; Fan et al., 2021; S. Kumar et al., 2017). According to Statista (2024), since 2012 the number of social media users worldwide increased from 1.48 billion to 4.76 billion in 2023.

Even though online environments have many benefits, they come with downsides too. Although some users try to make the online environment a safer and more usable place, there are users who still try to spread antisocial behaviours (S. Kumar et al., 2017; Mintz, 2002; Seife, 2014). These behaviours have dramatically increased over the years, including, disrespectful tone, hate speech, cyber-bullying, and cyber-harassment (S. Kumar et al., 2017; Lowry et al., 2017; Sheth et al., 2022). All these terms can be summarised as toxic behaviour in the online environment (Lowry et al., 2017; Messerschmidt et al., 2023). Users who experience toxic comments can also develop negative emotions or mental health problems. Examples of those side effects could be depression, anxiety, isolation, emotional harm, and increased suicidal ideation (Chick, 2020; Hsueh et al., 2015; Koutamanis et al., 2015; M. J. Lee & Chun, 2016; Rösner et al., 2016; Zaheri et al., 2020). In the literature, it could be found that cyber-bullying and online harassment can be caused by the factor of anonymity (Barlett et al., 2016; Lowry et al., 2017; Peebles, 2014; Slonje et al., 2012), while trolling can be affected by the mood and discussion context (S. Kumar et al., 2017).

On the one side, the design and structure of an online environment have been less researched in combination with toxic commenting behaviour. On the other side, the environmental design of offline environments has been researched in depth. Here, an appropriate theory has been developed to be applied to offline environments, which is called Broken Windows Theory (BWT). The BWT states that if disorders like graffiti, broken windows or vandalism are available in offline environments, it could influence an individual towards inappropriate behaviour (Ellis et al., 2020; Keizer et al., 2008; O'Brien et al., 2019b). In other words, such a disorder signalises that negative attitudes are seen as social norms in this environment (Ellis et al., 2020; Keizer et al., 2008; O'Brien et al., 2019b). According to Messerschmidt et al. (2023), this theory, applied in the online environment, would suggest that an unstructured and disordered designed environment indicates to the users that antisocial behaviour such as toxic commenting is seen as a norm or even leads the user to behave in a toxic manner.

1.2 Objective

This thesis aims to apply the idea of the BWT to an online environment by exploring whether the design of an online environment can influence toxic commenting behaviour. This has been done by analysing the commenting behaviour of users in an online environment where a redesign from disordered to less disordered design happened, and by answering the research question:

Does perceived disorder in an online environment increase toxicity?

1.3 Outline

To study the research question this work is structured as follows: The next chapter introduces the necessary background for this thesis, which is divided into BWT in offline and online environments. Additionally, an overview of the previous work of the BTW in offline and online environments is discussed. In Chapter 3, the applied methods are introduced and explained. This includes how the online environment has been selected and how the data from the online environment was collected, pre-processed, and analysed. In Chapter 4, the results are illustrated, while in Chapter 5 those results are discussed. In Chapter 6 the thesis is summarised.

2 Theoretical Background

2.1 Broken Windows Theory in an Offline Environment

2.1.1 Origin of the Broken Windows Theory

In March 1982, the social psychologists James Q. Wilson and George Kelling published the idea of the Broken Windows Theory (BWT) in the US American magazine "The Atlantic Monthly" (Kelling et al., 1982). This theory is related to criminology and tries to explain the neighbourhood's rising crime rate in the USA because of its environmental disorder condition (Kelling et al., 1982; Welsh et al., 2015). The Stanford psychologist Zimbardo (1973) conducted an experiment that served as the basis and source for the BWT, in which he investigated the vandalism that occurred when abandoned cars were placed in various neighbourhoods. Zimbardo (1973) discovered that the cars were more quickly dismantled for components and vandalised when left in high-crime disordered neighbourhoods instead of low-crime less disordered neighbourhoods.

The theory states that physical disorder (e.g., broken windows, graffiti, litter) and social disorder (e.g., vandalism, antisocial activities) can influence behaviour and attitudes, which can grow into more disorder, crime, and other incongruous behaviour (Ellis et al., 2020; Keizer et al., 2008; O'Brien et al., 2019b). In this case, a single disorder like the metaphorical "broken window" indicates that in this environment, negative attitudes are seen as social norms and will be tolerated in time, which triggers more incivility and crime (Ellis et al., 2020; Gau & Pratt, 2010; W. G. Skogan, 1992). Kelling et al. (1982) stated that "if a window in a building is broken and is left unrepaired, all the other windows will soon be broken". That is to say, a chain reaction of community decline will arise if a single instance of disorder is not fixed immediately (W. G. Skogan, 1992). This criminal behaviour will be spread further while the quality of life of the inhabitants declines (Gau & Pratt, 2010). The authors of this theory were not the first ones who discovered the phenomenon that disorders affect neighbourhoods. However, they were the first ones to claim that disorders could cause crime and attribute it to a causal role of disorder (Gau & Pratt, 2010; O'Brien et al., 2019b). According to Kelling et al. (1982) and W. Skogan (2015), disorder signals people from outside that this environment is not well controlled. As an outcome, crime can be practised without any consequence and attracts more criminals from the outside, while urban decay happens (Maskaly & Boggess, 2014; W. Skogan, 2015). For example, it signalises that the community or the police is either unable or unwilling to stop threatening behaviour and that informal social controls have collapsed (Gau et al., 2014; Maskaly & Boggess, 2014). Another outcome is that fear appears on the side of the residents, which impedes their ability to do something against incivilities and crime (Kelling et al., 1982). This indicates that if the neighbourhood can not handle "petty criminals" like beggars harassing other people on the street, a potential offender would imagine it as unlikely that they will call the police to identify or intervene in a robbery (Kelling et al., 1982; O'Brien & Sampson, 2015).

The BWT describes two pathways in which disorder leads to crime. The first is the direct way, where a broken window or other forms of uncontrolled disorder is a signal that lawbreaking will not be

punished, which in turn encourages others to break the law (O'Brien et al., 2019a, 2019b). The second way is the indirect way, which is associated with people who have no desire to engage in committing crimes (O'Brien et al., 2019a, 2019b). In this case, disorder illustrates an uncontrolled environment, where the danger is leading the community members to retreat from the streets or move out of the district (Kelling et al., 1982). This process leads to a reduction of people who would be able to deter uncivil behaviour, while crime increases (Bursik & Grasmick, 1993; Sampson et al., 1997; Shaw & McKay, 1942). In summary, Kelling et al. (1982) have two hypotheses, where the first is an "invitation to criminals", and the second hypothesis, where physical disorder causes fear and allows for social disorder and violence (Plank et al., 2009).

As an outcome, Kelling et al. (1982) suggest a strategy for the police, for handling crime explainable by the BWT. Kelling et al. (1982) recommend reducing stricter of small forms of disorder in order to handle and prevent the resulting crime from spreading out more. In the 1990s, Bratton - the commissioner of the New York City police department - had the goal to apply the idea of this theory in a practical way by focusing on a new policing strategy, which includes arresting individuals for "petty crime" (Gau & Pratt, 2008; Keizer et al., 2008). Also, the police is dealing harder with individuals in the neighbourhood who do not break any rule or law, but are seen as unpleasant in this environment, like drunks, panhandlers, prostitutes, and groups of rowdies (Maskaly & Boggess, 2014). Additionally, cleaning interventions like removing graffiti in the subways have been done and fare evasion in public transport has been punished harder (Gau & Pratt, 2008). This strategy is also called "broken windows", "zero-tolerance" or "quality of life" policing (Kelling & Coles, 1997; Silverman, 1999). This policing had also shown an effect in such a way that in New York the crime rate decreased dramatically between 1990-1999 (Corman & Mocan, 2005; DiIulio Jr, 1995; Kelling & Bratton, 1997).

2.1.2 Definition of Disorder in an Offline Environment

In the literature, the term disorder has been widely discussed and was agreed to be defined as a representation of a "minor violation of social norms" (Yang, 2014). According to W. G. Skogan (1990), disorder is something that deviates from the social norm, which determines how individuals should behave in a community. Based on Gau and Pratt (2010), many types of disorders are considered to be crimes. Scholars have defined a differentiation between two types of disorders called physical and social disorders. Physical disorder is related to how the environment looks like while in social disorder individuals are involved (W. G. Skogan, 1992). As stated by Sampson and Raudenbush (1999), in social disorder, neighbourhood members see inappropriate behaviour from strangers and see that as potentially menacing. On the contrary, physical disorder is related to the degradation of the environment (Sampson & Raudenbush, 1999; W. Skogan, 2015).

A physical disorder in a neighbourhood can be noisy, dirty, and run-down buildings which are in disrepair and abandoned (Ellis et al., 2020; Ross & Mirowsky, 2009). Some physical disorders just violate safety regulations like littering (W. Skogan, 2015). Other examples depict misdemeanours such as vandalism and graffiti (W. Skogan, 2015; Ellis et al., 2020). Most of them are just outcomes of lack of investment in maintenance (W. Skogan, 2015), whereas others like abandoned buildings

are related to economic collapse and leaving "dog excrement" is just a lack of regard for others (W. Skogan, 2015). Examples of social disorder in the neighbourhoods can be people who are drinking on the streets or taking drugs (Ellis et al., 2020; Ross & Mirowsky, 2009). A few social disorders, such as street preaching, are constitutionally protected behaviours. However, some activities like fighting are breaking civil rules (W. Skogan, 2015). Additionally, several common activities related to homelessness, like panhandling, sleeping rough, and searching dumpsters for food, are still legal in the USA (W. Skogan, 2015). Table 1 demonstrates further examples of social and physical disorders taken from W. Skogan (2015).

Social Disorder	Physical Disorder
Dumpster divers in search of food	Burned, abandoned, or boarded-up buildings
Street harassment/Cat-calling women	Broken streetlights
Recreational violence in pubs and clubs	Overgrown trees and shrubs
Excessive noise	Garbage strewn alleys
Loitering	Empty beer bottles visible in the street
Open gambling	Rats in the alley
Truant youth	Condoms/Needles on sidewalk
Street drug sales	Graffiti
Street prostitution	Broken windows
Public urination	Dog excrement

Table 1: **Overview of social and physical disorders (W. Skogan, 2015).**

2.1.3 Previous Work in Offline Environment

The BWT has been the subject of many studies since its publication. Most papers have focused on BWT policing in the USA, specifically in New York City. Nevertheless, it has been discovered that the BWT also works for other countries, such as Mexico. Based on the new work (especially the one done for Mexico), the researchers recommended new opportunities for police to prevent crime (Vilalta et al., 2020).

Even though this theory has been mainly focused on studying its impact on criminal research, its impact was extended to other areas, such as its context in offline environments. For example, in social psychology (Sampson & Raudenbush, 2004), law (Harcourt & Ludwig, 2006), as well as in public health (D. Cohen et al., 2000). In the area of public health, it was discovered that disorder in a neighbourhood has a connection to the infection of sexual diseases, leading to unsafe sexual health behaviour (D. Cohen et al., 2000). At the same time, other researchers explain that disorders may promote the consumption of alcohol and drugs (Browning et al., 2013; Rachele et al., 2016).

Another prominent study has been done in public in the town of Groningen in the Netherlands, in which the effect of the BWT on six different controlled field experiments has been tested (Keizer et al., 2008). Their study tried to simulate a disordered environment in some areas of Groningen to find out if it leads to crime or antisocial behaviour. The outcome was that individuals who noticed others violating social norms were likelier to do the same, which spreads more disorder in this environment (Keizer et al., 2008).

The BWT was also researched in areas covering tourist places (J. Liu et al., 2019). J. Liu et al. (2019) found out that a dirty environment signalises the visiting tourists that environment-damaging behaviours are allowed in this area. In the same disordered environment, pro-environmental behaviours will be ignored or even not encouraged. In simpler terms, if the tourist destination environment is clean without disorders, the tourists try to engage in a clean environment instead of ignoring it or even making it worse (J. Liu et al., 2019).

The BWT has also been studied in enclosed settings like in a hospital (Churruca et al., 2018; Ellis et al., 2020), in a school (Plank et al., 2009), or in a more private enclosed setting, which in this case was in the academic workplace in the department common room (Ramos & Torgler, 2012). To illustrate the latter case, the behaviour of academics in a clean and disordered environment was studied by Ramos and Torgler (2012): This controlled field experiment involved a homogeneous group of people. Although one might think that academic societies, commonly known as being the most educated ones, might behave in accordance with the rules in the context of littering, the opposite was demonstrated. The research found out that the BWT also works in cases where an indicator of disorder in a room leads to an increase of disorder or, in this case, littering. To put it differently, if an academic sees another academic breaking the social norms in a clean room, there is a chance that this academic will also litter in this same room.

2.1.4 Criticisms and Side Effects of the Broken Windows Theory

Despite being a well-known theory, BWT has received much criticism. One reason is that this theory was not empirically tested by the authors (Gau & Pratt, 2008; Harcourt, 2005; Link et al., 2017). This implies that without undergoing a thorough scholarly assessment, the BWT was swiftly converted from a publication into policy (O'Brien et al., 2019b). Even after several years, a number of scientific papers highlight the lack of validity of this theory (O'Brien et al., 2019b). Moreover, Kelling et al. (1982) never conducted a methodical investigation into the correlation between disorderly neighbourhoods and the crime rate (Young, 2013). On the contrary, some studies support the BWT (Savolainen, 2007; W. G. Skogan, 1992; Xu et al., 2005). According to some researchers, social conditions of the environment, as opposed to disorder, are a more reliable predictor of criminal risk (Taylor, 2001). For example, O'Brien and Sampson (2015) discovered that violent crime is more strongly correlated with social consistency failure than with physical disorder.

In another study showing disprove to the BWT, the researchers Sampson and Raudenbush (1999) used trained observers to investigate the streets in 196 districts in Chicago, where 15,141 streets are defined as disordered. The researchers observed that the association between recorded disorder and crime rates vanished when structural neighbourhood conditions such as poverty were taken into account (Sampson & Raudenbush, 1999). They claim that a third element that helps to explain the crime rate is solidarity among the populace together with a common expectation of public social control (Sampson & Raudenbush, 1999).

Also, the policing strategy inspired by the BWT has been criticised, especially at the time when it was used in the 90s in New York (Maskaly & Boggess, 2014). Few studies could confirm that broken windows policing could reduce crime in New York (Cerdá et al., 2010; Corman & Mocan,

2005; Kelling & Sousa, 2001; Rosenfeld et al., 2007). However, other studies were not able to prove this (Chauhan et al., 2011; Eck & Maguire, 2000; Harcourt & Ludwig, 2006; Joanes, 1999). The opponents of the BWT argue that the broken windows policing was not the reason why the crime dropped. They argued that the decline of the crack epidemic and that crime was decreasing in New York City before the police used their policing strategies. In other large cities, however, the crime rate went down without the aggressive policing strategy (Blumstein, 1995; Bowling, 1999; Eck & Maguire, 2000; Harcourt & Ludwig, 2006), or the economic upturn (Sampson & Raudenbush, 2001). Another big side effect is that this strategy could lead to racial bias (Harcourt, 2005; Taylor, 2001). As the police have a lot of discretion to distinguish who is disorderly, they can freely criminalise or judge communities of colour and groups who are financially disadvantaged (Collings-Wells, 2022; Roberts, 1998).

2.2 Broken Windows Theory in an Online Environments

2.2.1 Definition of Online Environments

In the literature, there is no clear definition of what an online environment is. According to Zhelev and Iliev (2023), an online environment is "*A place which can be reached via the internet or a local network which contains various information*". With this respect, online environments can be diverse. For example, they can be newspaper websites where a commenting section is available to be used by users. They may also be online forums created by online users, where certain topics are discussed. Furthermore, they can be Information Systems assets, such as websites or public-facing networks and systems of organisations. In this thesis, an online environment is defined as an area which can be reached via the Internet by using information and communication technologies. In this environment, communication takes place through text, images, emojis and emoticons, while this thesis focuses on text-based communication. In other words, commenting systems should be available on such platforms. In comparison to offline environments, in the online environment, individuals are more anonymous (Jaidka et al., 2022). Also, online environments are massive interactive media places that provide chances for more public discussions (Sobieraj & Berry, 2011).

One of the most used classes of online environments is Social Network Sites (SNS) (Dang, 2021). SNSs can be defined as Internet-based mass personal communication applications, where the site allows for the creation and exchange of user-generated content (Flury, 2017; Kaplan & Haenlein, 2010). These platforms allow the user to interact or self-present in a synchronous or asynchronous way to an audience (Carr & Hayes, 2015; Fernando et al., 2023). According to Aljafari (2019), a SNS must enable users to create profiles, connect with other users, engage in conversations, and upload content in real-time. A specific type of SNS is microblogging (Fernando et al., 2023). On that type of site, users share information with their followers in a very brief text message, which is mostly limited to 140 characters (Fernando et al., 2023; K. A. Mills & Chandra, 2011). Examples of microblogging sites are X (formerly known as Twitter) and Weibo (Ding & Qiu, 2017; Fernando et al., 2023).

A special characteristic of online environments, especially SNSs, is the highly arising toxic behaviour (Messerschmidt et al., 2023). One reason could be that anonymity gives users the freedom to communicate in an unhealthy way (Barlett et al., 2016). Toxic behaviour has also been studied on several platforms like X (Chatzakou et al., 2017; Davidson et al., 2017), YouTube (Dinakar et al., 2011; Y. Chen et al., 2012) or Reddit (Massanari, 2017; Messerschmidt et al., 2023).

2.2.2 Definition of Online Disorder and their Possible Side Effects

As with online environments, there is also no proper definition of online disorder in the literature. In this thesis, the physical disorder will be related to the website's visual appearance or platform design. As stated by Grimes et al. (2014), a poorly designed website (in terms of quality, credibility, and care of the website) conveys to the user that acting insecurely is a social norm on that website. Compared to the offline environments, where broken windows and graffiti signalise that disorderly behaviour is allowed in this neighbourhood, the same can be transferred and observed in the online environment (Grimes et al., 2014). This demonstrates that a low-quality website design discourages online users from behaving securely or with respect, for example. In contrast, a high-visual website signalises trust and indicates that the norm in this environment is secure (Grimes et al., 2014).

Similarly to disorder in offline environments, online disorder can also be diverse. Examples of online disorders can be an unprofessional asymmetric website layout (Bauerly & Liu, 2008) or the usage of low-contrast colours such as yellow, grey and white (Bonnardel et al., 2011; Bottomley & Doyle, 2006; Cyr et al., 2010). Moreover, high complexity and low-symmetry components on a website can make the website less aesthetically appealing (Bauerly & Liu, 2008). Other examples of online disorders are broken links on the website page, false or outdated "Contact and About Us" information, and broken image links (Abbasi et al., 2010; Y. Lee & Kozar, 2006). As stated by Fogg (2003), a broken link on a website can be interpreted by the user as an indicator that this website has been carelessly maintained or poorly designed. Abbasi et al. (2010) suggest that incorrect grammar, misspellings or complex URLs with a small number of workable links are indicators of fraudulent sites. According to Alsudani and Casey (2009), contrast dominance and image dominance can increase reliability and trustworthiness, but these prominent symbols have to be positively interpreted (Hynes, 2009; Lowry et al., 2014). Fogg (2003) found out that websites with good design will be evaluated as more authentic. To put it simply, navigation, visual and information design are all positively correlated to trust and satisfaction (Cyr, 2008).

Such online disorders can also have harmful side effects. An unclean website design can indicate to cybercriminals that this website has cybersecurity deficient, which encourages them to perform hacking attacks (Zadig & Tejay, 2010). In contrast, less disorder can indicate that this website has high Information Systems security (Zadig & Tejay, 2010). In the commenting behaviour, this could also have an effect since users may be dragged into deep conversations in the online environment. Aragón et al. (2017) found that changing a commenting section from linear to hierarchical can lead to intensive and meaningful discussions related to higher argumentation. It is also noted that a disordered online environment can also lead to more toxic or negative commenting behaviour (Messerschmidt et al., 2023). It is hard to explicitly define whether a particular website is considered

an online disorder since the features discussed above (colours or commenting sections) vary from user to user. Some users may consider a website to be disordered, while others do not.

In this thesis, the online disorder could have a similar effect as in the offline environment. Messerschmidt et al. (2023) hypothesise this too. This thesis assumes that perceived disorder in an online environment is a combination of the two pathways of the BWT as described in Section 2.1. In that case, it means that a website or platform design, which is structurally disordered and cluttered, could influence the user to perform more antisocial behaviour, which could lead to more toxicity (Messerschmidt et al., 2023).

2.2.3 Previous Research on the Broken Windows Theory in an Online Environment

In comparison to offline environments, where the BWT has been studied several times, there is less research in the context of online disorder in the form of a bad website design in online environments where BWT has been applied. Nevertheless, this section introduces a few studies that have evaluated the BWT in the online environment.

One of the first studies to evaluate BWT with online disorder is from Zadig and Tejay (2010). This study focused on using the BWT to understand how to prevent external hackers from attacking Information Systems from organisations. Their purpose was to find out if a reduction in disorder in the Information Systems, which heightens the illusion of security, would lessen external hacking attacks. To answer this question, they did a qualitative analysis of a case study. In this case study, the disorder in the Information Systems environment of the e-commerce organisation "TJX companies" has been studied. This company was a victim of a cyber-attack. As for data sources, they used secondary data such as news articles, court filings, legal documents, and other published information regarding the attack. They first analysed the case from the view of the internal "TJX" employees and then from the view of the external regulatory agencies. As a result of their analysis, they found out that online disorders such as insecure wireless protocols and lack of firewalls made the company an easy target of cyber-attacks. Zadig and Tejay (2010) suggest that if the company had reduced these disruptions, even if other security problems within their internal networks had not been fixed, the illusion of strong cybersecurity through the visual appearance of control would have been created and may have prevented the attack. When a hacker sees a less disordered online environment, the hacker is then demotivated to hack it since this person would think it might be time-consuming to hack that particular environment and thus might prefer to look for another victim (Zadig & Tejay, 2010).

Another study conducted in the area of information security focused on studying whether there is a connection between the BWT and poor website designs (Grimes et al., 2014). It discussed that users on low-quality websites would exhibit insecure behaviour in that online environment. Grimes et al. (2014) did a laboratory experiment by designing two websites with different quality while the participants had to create an account and a password on each website. The first website looks way more disordered, while the second website looks less disordered and more professional. The aim was to find if the participants were using insecure passwords on the disordered website in comparison to the less disordered website. To discover the secure behaviour of the users, they calculated the

password entropy. Their outcome indicated that a website with better quality and less disorder might lead users to more secure user behaviour on a website (Grimes et al., 2014).

The next study was conducted in the area of communication behaviour on a Spanish social news website called Menéame (Aragón et al., 2017). In comparison to the previous two studies, this study has not been combined with the BWT. Their aim was to discover how the commenting behaviour depends on changes in the conversation view by changing the conversation view: From a linear view into a hierarchical view, which is a tree structure of discussion threads. To that end, they applied graph network methods and discovered that in a hierarchical conversation view, the conversation was deeper, and the deliberation was higher in comparison to a linear conversation view (Aragón et al., 2017).

A similar study has been done by Fredheim et al. (2015) in the context of the BWT. The aim was to discover how commenting behaviour changed on the "Huffington Post" after an online commenting policy adjustment was applied. In this change, the users are not anonymous anymore. In their research, they made a "before" and "after" comparison by using different methods, such as time series analysis, to determine to what extent the quality of the comments has changed. Nevertheless, their focus was not on design-related but more on the anonymity of the users. A "broken windows" effect was found as a result of their study, in which comment quality improved even when the interaction with the "trolls" and "spammers" was excluded. Additionally, the number of comments decreased, yet the comment length increased (Fredheim et al., 2015).

One recent study on the BWT in online environments was done by Messerschmidt et al. (2023) focusing on commenting behaviour on SNS. They aimed to determine whether perceived disorder increases toxic commenting behaviour in SNSs. To this end, Reddit, an SNS online environment, was utilised to compare the comments of two corresponding subreddits with different design features. They collected the comments of these two subreddits and analysed the differences in toxicity by using Natural Language Processing (NLP) methods. In their study, the researchers took one subreddit pair, having the same discussion topics and similar descriptive metrics like community size or creation date, however, the visual design features were different. In the end, they collected comments from 2020-2022 from the same month for each subreddit, which are only text-based. The research discovered that the BWT can also work in an online environment. One finding was that the toxicity between 2020-2022 was consistently higher in the disordered environment in comparison to the less disordered subreddit. These results indicate that disorders in this online environment might lead the users to antisocial commenting behaviour by being toxic on that platform. Additionally, their analysis showed that the subreddit with less toxic behaviour attracts more users, while the average number of comments per user increased in the subreddit with more toxic commenting behaviour (Messerschmidt et al., 2023).

2.2.4 Research Gap and Developed Hypotheses

After an overview of the previous research in the area of BWT has been provided, it is noticeable that not much exploration has been done in this field. To the best of our knowledge, the only existing research in the field of BWT, as well as online commenting behaviour regarding the design

perspective, are from Messerschmidt et al. (2023) and Aragón et al. (2017). Meanwhile, the study from Fredheim et al. (2015) applied the BWT in online environments by comparing comments "before" and "after" a policy change without focusing on the design perspective itself. However, the analysis of Aragón et al. (2017) includes a comparison of a "before" and "after" change, which was only based on network analysis and not on NLP. Also, the studied comments were in Spanish and not in English. Messerschmidt et al. (2023) had a major focus on the comments themselves, which were inspected by applying NLP methods such as sentiment or toxicity analyses. Nevertheless, they made a comparison of two subreddits (from 2020 to 2022) with different design components that existed at the same time. In other words, Messerschmidt et al. (2023) did not focus on a design change of the complete online environment.

Although Messerschmidt et al. (2023) and Aragón et al. (2017) were able to find meaningful insights in this area, there are still a lot of potential open and research gaps that can be considered. For instance, an evaluation of a design change by analysing comments over time from the disordered to the less disordered platform using NLP methods has never been made. Another unexplored area is the behaviour of individual users, which was not considered in detail in any of the research. To put it differently, there was no user-based analysis of the commenting behaviour by focusing on the users who were active in this whole process over the entire period from the redesign from disordered to less disordered. Additionally, in the context of Reddit, there was no focus on the submission itself, which might also give insight into the BWT in an online environment.

Therefore, this thesis aims to extend the research of Messerschmidt et al. (2023), Aragón et al. (2017) and Fredheim et al. (2015) by finding out how the design component in an online environment can influence the commenting behaviour through the lens of the BWT. Primarily, the aim is to analyse the commenting behaviour in an online environment, where a design change from disordered to less disordered happened. The idea is to investigate and compare how the commenting behaviour of users has changed "after" the design change. In this case, Reddit was chosen as the online environment (Section 3.2). In comparison to Messerschmidt et al. (2023), for this thesis, Reddit is seen as the whole online environment, where a certain redesign happened from disordered to less disordered. Another extension to the previous work is the more extended time period. In this case, the comments between 2016 and 2022 have been explored, whereas the time range of the comments from the study of Messerschmidt et al. (2023) was between 2020 and 2022. Under these circumstances, the time range needed not to be affected by major events, such as the "Coronavirus pandemic" that took place from 2019-2022, which could impact the commenting behaviour (Henwood et al., 2023). To get more robustness for this research, two subreddits were selected that must deal with very similar topics which have been compared and analysed. The amount of collected data is much larger than in previous research. Additionally, submission-based analysis has been done by employing a subreddit crossline-based comparison of the commenting behaviour on submissions posted by the same user in both subreddits. *Overall, this approach can be seen as a comparison of two different districts (subreddits) of one city (Reddit), where the BWT has been applied (design change) by reducing the disorder*. At the same time, the individuals are very diverse for each community (subreddit). Since there are more possibilities to design the subreddits

"after" the redesign, all subreddits on Reddit no longer look visually the same and they can also look different by design. This makes it possible to follow a similar research task as Messerschmidt et al. (2023) did. In other words, it can be studied which subreddit appears to be more disordered "after" the redesign compared to the less disordered subreddit and if this impacts the commenting behaviour. Another extension is that the analyses have been done from four different viewpoints: Descriptive, content-based, user-based, and submission-based analyses. Especially the individual user-based and submission-based viewpoints have not been considered in any previous research. For the first two viewpoints, time series analyses are also carried out to get a better understanding of how the commenting behaviour has changed over time, which was inspired by the work from Fredheim et al. (2015). All the analyses of the comments are computer-aided with data science methods such as NLP methods or statistics (more detail in Section 3).

Based on those different analysis perspectives, four hypotheses were developed:

H1: The toxic commenting behaviour of users in an online environment (Reddit) will improve after a design change from disordered to less disordered.

H2: The number of active and unique toxic users will decrease after a design change from disordered to less disordered.

H3: All users, who were permanently active within a defined time range, will decrease their toxic behaviour after a redesign.

H4: The subreddit community that appears to be more disordered after the redesign will reflect higher toxic behaviour in comparison to the subreddit which is less disordered.

3 Methodology

3.1 Workflow

Figure 1 illustrates the workflow that has been done to answer the research question of this thesis. This workflow is divided into three phases consisting of several steps. The first phase illustrates how the online environment search has been done. The remaining steps show how the data from this online environment have been accessed, pre-processed and analysed to answer the research question. A detailed description and explanation of each phase follow in the next sections.

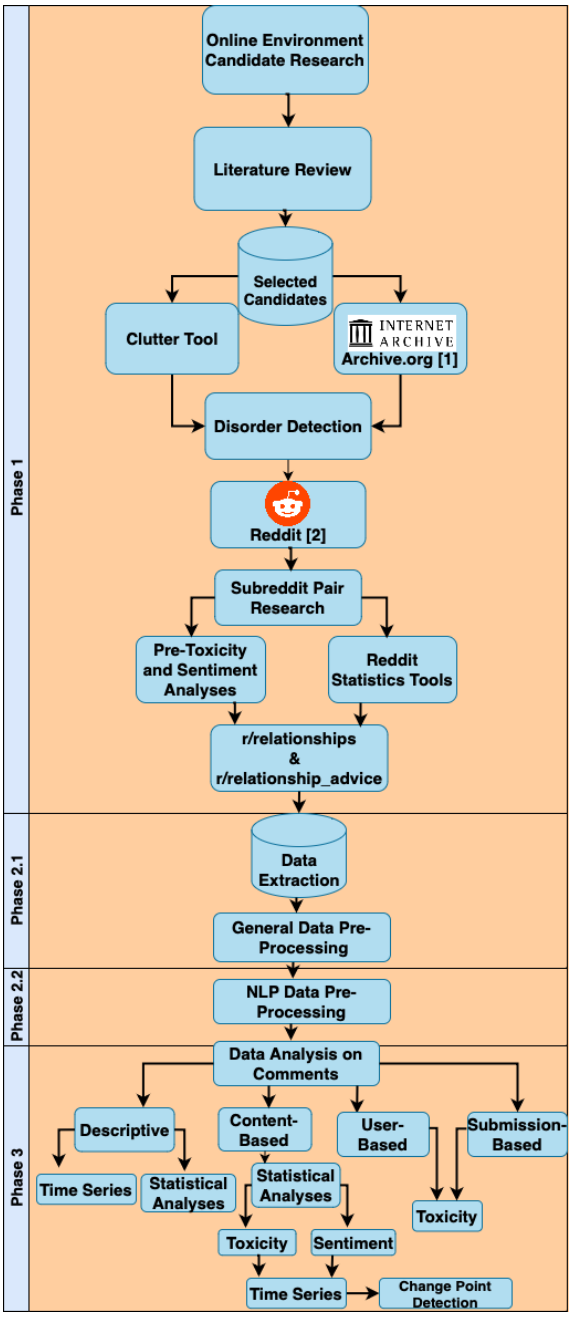

Figure 1: **Flow chart illustration of the workflow** [1-2] (Archive.org, 2024; Icon-Icons, 2024).

3.2 Phase 1: Online Environment Research

To study the commenting behaviour "before" and "after" a design change, a specific online environment is required. "Before" the design change, this environment should be disordered, and "after" the design change, the environment should be less disordered. Reddit was selected as a platform to complete this study. This section describes how this online platform was selected and briefly introduces Reddit. The selection process of the platform and subreddit pair research was inspired by the work from Messerschmidt et al. (2023), yet, with different hypotheses.

3.2.1 Online Platform Research Process

To determine the most suitable online environment for this research, pre-defined criteria that must be fulfilled were developed. First, the online environment must allow human interaction, which, in this case, is achieved through user comments. Second, the platform must have undergone a drastic design change in the past, as this thesis focuses mainly on environmental decay through design-related features. This design or structure change must be from a disordered online environment to a less disordered environment. To clarify, it is a comparison between a disordered design ("before" the design change) against a less disordered design ("after" the change). It is important to note that all other aspects should remain the same, except for the visual appearance of the online environment. Additionally, it is essential to legally access comments using an Application Programming Interface (API) or scraping tools. For identifying the online environment candidates, a literature review with various approaches has been employed in the selection process. In the first place, online environments have been utilised, which were applied in the research of BWT online area. Furthermore, any online environment that was referenced or utilised in the research field of online commenting behaviour was also included. The candidate platforms were selected according to the specified criteria mentioned above. In total, 145 candidates from online environments such as blogs, online newspapers, social media platforms, and Q&A pages like Stack Exchange or Quora were selected for the next step.

3.2.2 Disorder Design Detection

To verify whether the design-related criteria for the "before" versus ”after” comparison are congruous, the American digital library "Internet Archive" (archive.org) has been used to compare the design changes. Internet Archive (1996) is an online public library, which provides public-domain digital material like internet sites and other cultural artefacts in digital form (Burnett, 2009; Internet Archive, 1996). Through Internet Archive (1996) it is possible to get access to web history through many years. By accessing the history of a desired website, e.g. by traversing all available snapshots of the online environments, it is possible to determine if a platform was disordered "before" a change and less disordered "after" this change.

As mentioned in Section 2.2.2, it is not easy to evaluate by eye whether an online environment is disordered or less disordered, as this is very subjective for every individual. To bring an objective factor into this evaluation selection process, an automatic detection approach to the degree of

disorder is needed. To solve this, the clutter score has been calculated, which gives insights into how disordered a website design is (Rosenholtz et al., 2007). For this, the Python implementation published by Kargaran (2021) has been applied. It allows for calculating clutter scores for snapshots or image files of the website design (Kargaran, 2021). This tool calculates two scores that give a hint of whether an online environment is disordered or less disordered. The first score is the Feature Congestion (FC) score. It is a measurement that is based on three key features: Colour variability, luminance contrast and orientation contrast. The second score is the Subband Entropy (SE), which gives the same characteristics as the FC, except for the orientation variance. More precisely, it tries to group similar objects in an image. The values for both measurements are between 0 and 10. The scores can be interpreted by the following concept: The higher the score is, the more disordered the design of the online environment is (Hammami & Afram, 2022; Kargaran, 2021; Rosenholtz et al., 2007). After all the steps of this selection process were completed, the results showed that Reddit is an ideal online environment for studying the research question (see Section 1.2).

To summarise, Reddit fits every pre-defined criterion. First, on this platform, human interaction is possible by using comments. Second, the platform data is publicly available and can be scraped for free with different tools. The third, which is the most important criterion, is the design change from a disordered to a less disordered platform. To be more precise: In April 2018, Reddit had a major redesign, which was its first visual and structural change after many years (Liao, 2018; Loten, 2018; Pardes, 2018). The results of the clutter score for the complete platform can be seen in Table 2. The three rows show the FC and SE results for the front, submission and comment pages, respectively. With regards to Reddit's front page, the FC and SE scores "before" the change (2017) are 5.527 and 3.442, respectively. Whereas "after" the change, the FC score is 3.390, and the SE score is 2.896. This gives a slight hint that "before" the change, the Reddit environment was much cluttered and disordered compared to "after" this design change. Table 2 does not show clutter scores for submission and comment for 2018 because no images were found for the same submission and comment sections for that year. Similar score results could be seen by comparing the commenting section and the submission structure. The images of how the design occurred "before" and "after" the change can be found in Appendix A in Figures 18-24. In the following section, the online environment Reddit will be introduced.

Page	Before (2017)	After (2018)	Now (2023)
Front	FC:5.527/SE:3.255	FC:3.390/SE:2.896	FC:3.791/SE:2.948
Submission	FC:7.149/SE:3.480	-	FE:6.158/SE:3.052
Comment	FC:5.534/SE:3.089	-	FC:4.065/SE:2.438

Table 2: **Overview of clutter score for front, submission and commenting sections.**

3.2.3 Reddit as an Online Environment

Reddit is a public social media platform that was launched in 2005 (Trager et al., 2022). This platform can be seen as a social news website with a bulletin board system, where users can share news or other content (Fabian et al., 2015; Koh, 2018). According to Reddit (2020), there were more than 52 million active users and over 138,000 active communities called subreddits (Marotti,

2018; Proferes et al., 2021; Reddit, 2020). On this platform, more than 303 million submissions and 2 billion comments have been posted by users in 2020 (Reddit, 2020). According to Similarweb (2023), Reddit is the 18th most visited website in the world since October 2023. Reddit is a website whose content is mainly user-based. This content can be images, videos, links and texts, in which various topics are discussed (Nudd, 2014).

Figure 2, provided by Medvedev et al. (2019), illustrates how Reddit is structured. When entering the website, the top (front) page of Reddit is shown. This page is filled with the top-voted submissions from all subreddits that the registered user is following. Registered users can write a post/submission, which contains a title, an external link, a self-written piece of content, an image, or a video, which immediately becomes available to the whole audience of Reddit, who can vote and/or comment on it. A user also has the possibility to enter the front page of a specific subreddit directly, where all the submissions for this subreddit are available. Each submission and the included comments can be commented or voted on. According to Medvedev et al. (2019), the commenting structure can be viewed as a rooted tree by the reply-to relation to other comments or the post/submission itself. To put it differently, the first node is the post/submission itself, and the following nodes represent the corresponding comments. If there is a linkage between two nodes, it implies that this is a reply-to relationship between these comments.

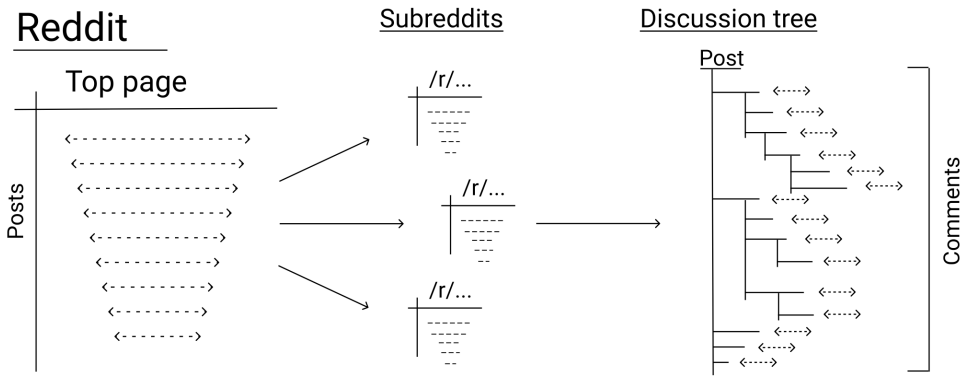

Figure 2: **Reddit page structure (Medvedev et al., 2019).**

Regarding users, there exist several possibilities that a registered user can perform on Reddit, for example, posting a submission, commenting on a submission or a comment, voting on a submission or a comment, and sharing the submission with other Reddit users (Medvedev et al., 2019; Silver, 2021). The voting system is also special in comparison to other social media platforms. Only registered users can upvote, which means giving a positive vote, or downvote, which means providing a negative vote on submissions and comments (Medvedev et al., 2019). In comparison to Facebook, YouTube or Instagram, where these platforms apply algorithms to direct attention to popular content, Reddit is applying its voting system. This system encourages submissions with higher upvotes to pop up on the front page of Reddit's top submissions (Messerschmidt et al., 2023). In accordance with Proferes et al. (2021) and Silver (2021), Reddit is a semi-anonymous social media platform, where users can choose any username they want, which could be used to hide their identity. Therefore it is challenging to obtain demographic information about Reddit

users. Nevertheless, any activity done by these users on Reddit is stored on their profile (Silver, 2021).

A special feature of Reddit is called subreddit. Subreddits can be seen as topic-based sub-communities that can co-exist (Hessel et al., 2016; Messerschmidt et al., 2023). On these subreddits, a discussion about a certain topic is done by the registered users. It should be noted that these topics can be diverse. Consequently, for any topic, there often exists a subreddit where users with the same interests can be members of that subreddit. A subreddit is defined with the prefix "r/" followed by the name of the created community (Reddit, 2023c). There are currently three main types of subreddits: Public, restricted, and private subreddits. Public subreddits are communities, in which each user can view comments and submissions. Restricted subreddits are subreddits, in which anyone can view and/or comment with the condition that moderators approve these members (Reddit, 2023d). Finally, private subreddits are communities, in which users need to request to enter and participate in the requested subreddit and the activities there (Reddit, 2023d). Most subreddits are public, where everyone with or without a Reddit account can view this content (Proferes et al., 2021). This thesis focuses mainly on the public sub-communities. Another special feature of Reddit is the design opportunities for each subreddit. Every subreddit could have a different design, including the use of various colours, shapes and backgrounds, which also include the submissions themselves. The submissions can vary in size, form, media type or visual appearance. Even if two subreddits have the same topic, the visual appearance can be completely different (Messerschmidt et al., 2023).

Anybody is allowed to create a subreddit. People who create such subreddits are called moderators. These are volunteers who are interested in the subreddits topic (Matias, 2019; Messerschmidt et al., 2023; Seering et al., 2019; Singh, 2019). Additionally, these moderators can also select more users to be moderators on this subreddit. Moderators have control of the created subreddit and are allowed to set the rules or do modifications to this subreddit. Such modifications include visual features like background, images, colours or post flair. Also, moderators are allowed to ban users from participating in the subreddit if they violate their rules (Singh, 2019). For instance, when users attack other users (Singh, 2019). To sum it up, every subreddit follows different rules defined by the moderators, which must be abode by the users (Messerschmidt et al., 2023).

As stated by Proferes et al. (2021), Reddit is also very popular for many controversial events. This includes the identification of the terror attack in Boston (Starbird et al., 2014), the massive leak of hacked celebrity photos (Marwick, 2017), the coordinated attempt to take on short sellers of GameStop stocks (Roose, 2021), or sometimes with racist (Mittos et al., 2020), sexist (Farrell et al., 2019), and vitriolic political discourse (R. A. Mills, 2018). This indicates how attractive the data of Reddit can be for researchers (Proferes et al., 2021), but also for answering the research question, as this platform offers many diverse data resources and had a design change that can be studied.

3.2.4 Subreddit Pair Research

Since Reddit is chosen to be studied as an online environment, it is now essential to decide which subreddit will be selected for the study. To enhance the robustness of this research, the decision was

made to select two subreddits for separate analysis and subsequent comparison. As described in the previous section, many subreddits exist covering diverse topics. Nevertheless, according to Kitchens et al. (2020) and Messerschmidt et al. (2023), different topics can lead to different degrees of toxicity, as individuals or groups can have extreme views, which may lead to a polarised discussion. For example, a subreddit focusing on political topics may have different toxicity degrees compared to a subreddit, which is an online book community (Messerschmidt et al., 2023).

As this thesis focuses on studying the perceived disorder "before" and "after" a design change in the context of toxicity in an online environment, especially by doing a submission-based analysis, selection criteria are again needed. These criteria assist in finding the right subreddit pair choice. Those criteria are also inspired by Messerschmidt et al. (2023). Since the subreddits were all very similar in the design perspective "before" the change in 2018, and the design configuration options were rather limited at that time, it did not matter how the subreddits were designed visually "before" this change.

The first criterion is that the topics of the subreddits must be the same or similar for a fair comparison. As described above, different topics can lead to a different degree of toxicity (Messerschmidt et al., 2023). This criterion is especially important when applying submission-based analysis by analysing how the user communities comment on the same submission on different subreddits. The second criterion is that the subreddit pair must be public to access the data. The third criterion is that both subreddits should have similar descriptive metrics at the beginning of the selected time range for analysis. The reason is that a newly created community can not be compared easily with an established one that exists longer, because it might have more users and comments. The subreddit metrics that are focused on are the community size, the comment and submission ratio, and the creation date of both subreddits. The fourth criterion is that the subreddit language is mainly English, and the comments should be text-based since the analysis is done with NLP methods (more details in Section 3.4). Another reason is to minimise external influences from third-party content (Messerschmidt et al., 2023). The last and most important criterion is that the subreddit pair should not be affected by external factors or events from offline environments. For example, a subreddit that focuses on politics is less suitable because certain events, such as the presidential elections, could influence the commenting behaviour (Hiaeshutter-Rice & Hawkins, 2022). In this research, however, the focus is on the influence of the design and structure of the website concerning commenting behaviour. Therefore, subreddit pairs to consider are the ones that do not directly manipulate commenting behaviour despite certain external factors. It was also taken into consideration that subreddits, which already show topics where toxic behaviour is the norm, should be avoided. For that reason, both subreddits should not be too toxic or negative.

To fulfil all the criteria, a preliminary analysis was carried out to identify the appropriate subreddit pair. In the first instance, subreddits that met all the criteria were selected and then analysed with a sample of 1,000 comments. To discover if the descriptive metrics of both subreddits are similar, the two statistical tools Anvaka (2020) and Subreddit-stats (2023) were used to check the descriptive criteria, as well as the overlap of users. To look over whether the subreddit pair is not too toxic or

negative, toxicity and sentiment analysis (described in Section 3.4), as well as word clouds, were employed for the comments with the Communalytic tool (Gruzd & Mai, 2022).

After this selective process, the final subreddit pair that has been selected is related to the topic relationships. The resulting subreddit names are r/relationships and r/relationship_advice. By choosing the topic of relationships, it was possible to find a topic that is discussed in everyday life, where the mood is unlikely to be manipulated by real-world events like political elections. The reasons why these subreddits have been chosen are the following: 1) Both share a similar user base and keyword usage. This assumes that the content which is discussed should be similar. 2) The two communities have been created in a similar time range (r/relationship_advice in 2009, whereas r/relationships in 2008). 3) As these subreddits have existed for more than 13 years, it signifies that both are long-lasting subreddits. 4) The pre-NLP analysis revealed for a small sample of comments that both subreddits are not too toxic or negative. Important is also the fact that r/relationships and r/relationship_advice have a high overlapping number of users relative to the community size across all investigated pairs, which makes a fair comparison and the submission-based analysis possible. In the Appendix C, the appearance of the "before" and "after" the change for each subreddit for the years 2017, 2018 and 2022 can be seen. In addition to that, the clutter scores for both subreddit front pages over the years can be found in Appendix B, which shows that r/relationship_advice is more cluttered compared to r/relationships.

3.3 Phase 2.1: Data Extraction, Pre-Processing and Description

Unfortunately, Reddit announced in April 2023 that using the Reddit API and third party is not free anymore (KeyserSosa, 2023). Consequently, Pushshift "*zstandard*" compressed "*ndjson*" files were used which were hosted on Academic Torrents (stuck_in_the_matrix et al., 2020). These files contain comments and submission data from Reddit. One advantage of why this is an adequate solution is the fact that these files are separated by submissions and comments for every top 20k subreddit. As both subreddits are included in this list, this method was suitable. Also, the data do not need to be scraped from the very beginning, which saves a lot of computational time. To get the data, firstly, the files with the comments needed to be downloaded for both subreddits. Later, the data was extracted and compressed into CSV files with a Python script available by Watchful1 (2023) in the corresponding GitHub repository.

Before all the comments can be extracted, a time range should be defined. It was decided to collect and analyse the comments between 2016 and 2022. The intention for this decision is based on the year (April 2018) when the design change was applied, and therefore it was important to have at least two years "before" the change to encourage a fair comparison. The motive for the choice to until 2022 and not 2023 is the fact that there was no data available for 2023 in the "*zstandard*" compressed "*ndjson*" files. Nevertheless, since both subreddits have nearly millions of comments just for one or two months, it was decided not to choose all comments between 2016 and 2022, as it would be too much data and would be beyond the scope of this thesis. Therefore, it was decided to take the data published on both subreddits from 1st to 30th November between 2016 and 2022. November was chosen because it was the month with the least number of comments overall than the other months.

Also, it was important to have a month "after" the design change month since the design change was not visible to everyone in April (McLaughlin, 2018). Therefore, it was essential to choose a month like November instead of a month that directly comes after April, e.g. May or June. In total, 5,464,725 comments have been collected (r/relationship: 1,475,264 and r/relationship_advise: 3,989,461). Additionally, for the submission-based analysis, all the submissions with the submission ID, author, and submission title, which have been posted on both subreddits by the same user, have been collected. In the Appendix D, the number of all comments for each year in November, separated by communities, is displayed.

In the next step, an initial or general data pre-processing was employed, to clean the data discussed above. First, comments have been removed, which had been wrongly formatted, and had wrong or NA values. This can be, for example, the wrong subreddit ID or the date when the comment was created does not match the defined time range. Additionally, the created date for the comment has to be converted to a day/month/year format, as the format of these values is UNIX timestamp. Also, all comments have been removed, which were not written in English, since the focus of this thesis is to analyse English comments only. The next step was to drop all the comments, which have the values "removed" or "deleted", instead of a meaningful text-based comment. Additionally, empty comments are also removed, as these rows can be summarised as NA without content. Moreover, all comments have been deleted, which only contain URLs, HTTPS links, subreddits mentioning's or symbols like "!","?" or "#". Furthermore, all comments with a length of 1 have been dropped out, as these comments do not add any further information for the analysis. In the cleaning step, duplicates were included in the analysis, because those comments could contain common phrases like "thank you". Deleted users were included too since their comments are still available. As the focus of this thesis is on human interaction through comments, comments generated by bots have been removed too. To do this, all comments, where the author's name was "AutoModerator" have been removed. Also, with the help of B0tRank (2023), which included a list of bots that exist on Reddit, all the other bot comments could be detected and removed. Nevertheless, a lot of manual pre-processing has been done for detecting the bot comments. There were bot comments that appeared at first glance very authentic. For that reason, it was still difficult to distinguish between user and bot comments. To this extent, a manual search has been done by looking at keywords like "bot", "blip", "blop" or other standard phrases, which bots write in a comment. Finally, all columns have been removed except for those that are listed in Table 3. After this initial pre-processing step, the final dataset included 3,763,505 comments (r/relationships: 1,200,679 and r/relationship_advice: 2,562,826).

Feature	Description
Author	Account name of the poster.
Body	Comment itself in string format.
Created_UTC	UNIX timestamp, that refers to the time of the comment creation.
Link_Id	Identification of the submission, where that comment is in.
Karma Score	Number of upvotes, minus the number of downvotes.
Subreddit	Name of the subreddit, that the comment was posted in.

Table 3: **Data description and corresponding features (Baumgartner et al., 2020).**

3.4 Natural Language Processing

The primary objective of this study is to analyse the commenting behaviour of users of a specific subreddit pair. Since the comments are text-based, a method that can process such text is required. Such an approach is called Natural Language Processing (NLP), which takes a text-based input and analyses it (K. B. Cohen, 2014; Vijayarani et al., 2015). This approach is used in different areas, for example, in Computer Science, Linguistics, Mathematics, Artificial Intelligence, Robotics, and Psychology (Jusoh & Al-Fawareh, 2007; Vijayarani et al., 2015), as well as in the research of commenting behaviour on social media platforms (Jusoh & Al-Fawareh, 2007; Vijayarani et al., 2015). In this section, two NLP-based models will be introduced, which have been applied in order to answer the research question (see Section 1), followed by a section on how the text-based comments have been pre-processed for those models.

3.4.1 Sentiment Analysis

The first NLP method that has been used to analyse the comments is sentiment analysis. According to B. Liu (2022), sentiment analysis can be defined as a field of study that analyses people's opinions, sentiments, attitudes, and emotions towards entities such as products, services, individuals, issues, events, topics and their attributes. This would be done by applying computational classification on a text to detect if it has a specific polarity (like positive, negative or neutral)(Hutto & Gilbert, 2014; Melton et al., 2021). This approach is widely practised in different research fields like sociology, marketing, psychology, economics, and political science (Hutto & Gilbert, 2014). One major reason for utilising this method is the fact that it is one of the most frequently used methods for analysing data on social media (Taboada, 2016; Yue et al., 2019). There is research where sentiment analysis is used for large social media platforms such as X and, in this case, Reddit (Brown & Coyne, 2018; Melton et al., 2021).

There are different ways how sentiment analysis can be done. One approach is the machine learning-based method, where a classifier is built that has been trained on sentiment-labelled data such as sentences or documents. With this supervised learning technique, the resulting model can distinguish a positive text from a negative one (Taboada, 2016). Another approach is the lexicon or dictionary-based method. The idea of this method is to follow specific rules, where the sentiment values of a text are determined from the sentiment orientation of each word in the text by using an existing dictionary (Taboada, 2016). This dictionary includes words with different polarities. If a new text is encountered, the words are matched to those in the dictionary and their values are then aggregated using various algorithms (Taboada, 2016). The semantic orientation for the entire text is produced by aggregating the positive and negative values of the words in the text (Taboada, 2016). In this thesis, the lexicon-based approach has been applied. The reason is that this method is robust in comparison to other approaches without even changing the dictionaries when considering different domains (Ortigosa et al., 2014; Taboada et al., 2011; Taboada, 2016). As stated by Brooke et al. (2009), this method is, in the context of a new domain, still more robust in comparison to labelling data for a classifier. Also, the lexicon-based model uses the

linguistic information contained in the text. Thus, it has a good synergy between computational and linguistic approaches (Taboada, 2016).

In the present work, the Valence Aware Dictionary and sEntiment Reasoner (VADER) have been used to determine the emotions of the Reddit comments and to classify them according to multi-class sentiments. VADER has been developed by Hutto and Gilbert (2014) and is a rule and lexicon-based sentiment analysis tool (Elbagir & Yang, 2019; Hutto & Gilbert, 2014). This tool has been applied because it is well-suited for social media data (Hutto & Gilbert, 2014). The authors also conducted a study by comparing VADER with 11 other sentiment analysis tools. Out of this comparison, the authors discovered that VADER improved the benefits of traditional sentiment lexicons (Hutto & Gilbert, 2014). This model provides as output for each comment a score being positive, negative, or neutral- sentiment in addition to a compound score. In this thesis, the compound score, which is mostly applied in research, is the central metric that measures the sentiment for each comment. The compound score is a single value, which is calculated by summing up the three scores (positive, negative and neutral) of each word in the lexicon, adjusted by the rules, and then normalised (Hutto & Gilbert, 2014). The output is a value between -1 (for comments being extremely negative) and +1 (for comments being extremely positive). With a given threshold, this score indicates and classifies a comment as positive, negative or neutral. In this thesis, the thresholds suggested by Hutto and Gilbert (2014) have been applied. If the compound score is ≥ 0.05, the comment will be classified as positive. If the score is $\leq$ -0.05, it will be classified as negative, while a neutral compound score will be classified by the condition if the compound score is $>$ -0.05 and $<$ 0.05 (Elbagir & Yang, 2019; Hutto & Gilbert, 2014). This approach has been applied to all comments in this thesis.

3.4.2 Toxicity Analysis

The second NLP method used in this thesis, which is also the main focus, is toxicity analysis. As defined by Lowry et al. (2017), toxicity is a deviant behaviour incongruent with societies dominating values and norms (Messerschmidt et al., 2023). Especially in the online environment, toxicity is a behaviour that violates social norms and can be defined in different forms, such as hate speech against groups or individuals, trolling or even cyber-bullying (Messerschmidt et al., 2023; Obadimu et al., 2019). The literature defines online toxicity as the usage of "rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion" (Märtens et al., 2015; Nobata et al., 2016; Wulczyn et al., 2017).

In this thesis, the Perspective API has been utilised to identify and classify whether a comment can be seen as toxic or not (Google Jigsaw, 2023h). This API is based on a trained machine learning model (Google Jigsaw, 2023h) and has been developed by Jigsaw and Google's Counter Abuse Technology team in a collaborative research initiative called Conversation-AI (Google Jigsaw, 2023a). One reason why this API has been applied is the fact that the developers of this API employed the previously mentioned definition of toxicity in their model (Google Jigsaw, 2023b; Messerschmidt et al., 2023). Another reason is that this model has been trained on online forum comments, which could also be a good fit for Reddit data (Google Jigsaw, 2023e). The trained models were multilingual Bidirectional Encoder Representations from Transformers (BERT) based

models that were trained in different languages on millions of comments from different online forums such as Wikipedia and The New York Times. Those trained models were integrated into single-language Convolutional Neural Networks (CNNs) for each supported language (Google Jigsaw, 2023c, 2023e). This tool might assist human moderators in detecting toxic content and thus make the online environment a safer place for users (Jain et al., 2018).

In academic research, many scholars use the Perspective API to study toxicity with social media data, especially on Reddit data (Gruzd et al., 2020, 2023; D. Kumar et al., 2023). Furthermore, this tool has been employed in previous works, where the BWT has been applied in the online environment (Messerschmidt et al., 2023). Perspective API is a free tool that is available to use with only two requirements. First, a Gmail account with access to the Google Cloud console is necessary, while the second requirement is the API key that must be requested from the API itself (Google Jigsaw, 2023g). The unique feature of Perspective API is that in addition to detecting *Toxicity*, it can also classify five other subcategories, namely *Severe Toxicity*, *Identity Attack*, *Insult*, *Profanity*, and *Threat*. Table 4 gives an overview of all categories with additional descriptions as defined by Google Jigsaw (2023b). Even if all the other attributes from Perspective API have the same or similar names, those will be seen as subcategories of the main definition of toxicity.

Attribute Name	Description
Toxicity	A rude, disrespectful or unreasonable comment that is likely to make people leave a discussion.
Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
Identity Attack	Negative or hateful comments targeting someone because of their identity.
Insult	Insulting, inflammatory or negative comment towards a person or a group of people.
Profanity	Swear words curse words, or other obscene or profane language.
Threat	Describes an intention to inflict pain, injury or violence against an individual or group.

Table 4: **Overview of Perspective API attributes and the corresponding description (Google Jigsaw, 2023b).**

To determine if a comment is toxic, every comment receives a probability score between 0 and 1 for each category. The higher the score, the more likely a reader would perceive that comment as containing a given attribute. Nevertheless, it does not mean that this comment is more toxic. For example, if a comment has a score of 0.7 for the category *Toxicity*, it means that 7 out of 10 people would perceive that comment as toxic, while a score of 0.9 means that 9 out of 10 readers perceive this comment as toxic (Google Jigsaw, 2023a, 2023d; Messerschmidt et al., 2023). To classify a comment as one of the toxicity attributes, a threshold is necessary. In this case, the threshold 0.7

or higher recommended by Perspective API for social science researchers has been used (Google Jigsaw, 2023d). This threshold has also been applied in other social science research areas, which makes this selection reasonable (Gil et al., 2019; Google Jigsaw, 2023a, 2023d; Hua et al., 2020; Messerschmidt et al., 2023; Nogara et al., 2023). Unfortunately, it can happen that a comment does not get any score and will be ignored by the API. After a preliminary analysis, it could be found that most of the comments that were not assigned to a score have a positive or neutral sentiment, and at the same time, those comments have an average comment length of 3 words. In other words, there is no content to classify this properly. To avoid a loss of data, it was decided to fill in each comment that was not assigned a score of 0.

Since there are more than 3 million samples, running Perspective API locally would have a high computational time. For that, it was decided not to run this analysis with a scripting language such as R or Python on a personal laptop. Therefore, for this thesis, the web tool Communalytic, developed and maintained by the Social Media Lab at Toronto Metropolitan University's Ted Rogers School of Management, has been applied (Gruzd et al., 2020). Communalytic defines itself as a social science research tool for studying online communities and discourse (Gruzd et al., 2020). This tool has different functionalities like collecting, analysing, and visualising publicly available data from different social media platforms (Gruzd et al., 2020). In this application, the Perspective API is integrated too. As described in Section 3.2.4, this tool has also been used for the pre-analysis. For academic studies, there is a free version available. Nevertheless, this version can only accept 30k comments and takes approximately 7-8 hours to provide the corresponding analysed results. For that reason, it was decided to split all the data into chunks of 30k samples and calculate the Perspective API score for each chunk individually. It is essential to mention that this approach has no disadvantages for the analysis. To save time, four accounts were created to reduce the waiting time for toxicity analyses.

3.4.3 Phase 2.2: NLP Data Pre-Processing

Since NLP methods are mainly based on Machine Learning models that require cleaned data, it is an essential step to clean the data through pre-processing methods (Khanaferov et al., 2014; Neutatz et al., 2021). This pre-processing step aims to bring the text into a format that makes it easily understandable, predictable, and analysable by the algorithms (Tabassum & Patil, 2020). For example, HTML links, symbols or abbreviations have no content but negatively influence NLP methods (Alsuliman et al., 2022; Hegazi et al., 2021). There exist various approaches to clean the data in this domain. For example, removing stop words, lemmatisation or stemming (Tabassum & Patil, 2020). In the VADER sentiment analysis, the authors state that pre-processing is unnecessary (Hutto & Gilbert, 2014; Thapa, 2022). According to them, sentiment heuristics play an important role in estimating the writer's mood because punctuation, capitalisation, and trigrams can amplify the mood (Hutto & Gilbert, 2014; Thapa, 2022). Nevertheless, no standardised pre-processing approach could be found in the literature for the Perspective API. Also, studies that utilised this approach did not clean or mention any pre-processing steps needed (Gruzd et al., 2020; Messerschmidt et al., 2023; Shen & Rosé, 2022). The developers of this tool also did not mention

any pre-processing steps. To find a solution, the Perspective API support team was contacted, who gave specific recommendations to the case of Reddit data (Google Jigsaw, 2023f). Some recommendations were handed out, which were perhaps not necessary for this analysis. Therefore, different approaches were tested in advance for a small sample of comments, and the toxicity score was calculated for each of the comments. The same data has been used for the different analysis approaches. The aim was to find a pre-processing method in which as many comments as possible should be assigned a Perspective API score. As a result, the following pre-processing steps have been chosen and applied to the data sets:

1. Use Reddit cleaner Python library
2. Remove HTML links or subreddit mentionings
3. Remove special characters
4. Convert emojis and emoticons into text
5. Write out shortcut words like "gf" ,"bf","OP"

First, the Python library Reddit Cleaner has been used (Leitner, 2020). With this tool, the markdown and LaTeX formatting that appears in the comments were removed. After that, the HTML links and subreddit mentions were removed, as these have no content. As for the third step, special characters have been removed. In the fourth step, emojis and emoticons have been converted to text. Since emojis and emoticons will not be recognised by the Perspective API, those were converted to text. For instance, instead of "👍", this emoji was converted into a "thumbs up" as text. Converting the emojis and emoticons was performed by Kim and Wurste (2023). In the last step, shortcuts or words like "gf", "bf", and "OP" were written out. It is important to mention that capitalisation or punctuation have not been cleaned, as those can show emotions, too. For example, if someone writes something in capital letters, such as "I DON'T LIKE THIS", it could mean that this user is angry (Savigny & Purwarianti, 2017). To make a fair comparison, the sentiment analysis has also been applied to the same data.

3.5 Change Point Detection

Since this study aims to analyse the commenting behaviour in the context of toxicity (or other features) "before" and "after" a design change over the years, a time series analysis has been conducted in this thesis. As defined by Esling and Agon (2012), a time series is a collection of values that are sequentially measured over a period of time. Time series analysis is well-known also in social media research (Fredheim et al., 2015; Pellert et al., 2022; Thelwall, 2014).

Generally, this time series analysis aimed to discover whether trends are available by observing the results as line plots over the years and how the design change could affect them. Nevertheless, it is difficult to only consider the plots as a final decision on whether a certain change happened. To bring significance to this analysis, a change point detection has been applied to these time series, which is also common in different research fields (Matteson & James, 2014). Change point detection can be defined as a problem in detecting drastic changes in the data when the property of the time series

changes (Aminikhanghahi & Cook, 2017). According to Sachdev (2023), a change point is a point in time where a significant change in statistical properties, such as mean or variance, has happened. This method is a well-established approach used in time series analysis (Basseville et al., 1993; Cabrieto et al., 2017). It is used in various domains such as Climate Science, Economy or Medicine (Cabrieto et al., 2017; J. Chen & Gupta, 2012). Also, the change point analysis has been applied to social media comments by analysing COVID-19 data concerning sentiment in X (Theocharopoulos et al., 2023), as well as on the same platform regarding financial events (Qu et al., 2016).

Change point detection can be done in different ways, including fully parametric approaches or distribution-free methods (non-parametric) (N. A. James & Matteson, 2013; Matteson & James, 2014). In the parametric approach, the distributions belong to some known family, where also the likelihood function is an important component (Matteson & James, 2014). The non-parametric approach focuses more on the estimation of density function or using rank statistics (Matteson & James, 2014; Kawahara & Sugiyama, 2012; Lung-Yut-Fong et al., 2015).

In this work, the non-parametric, hierarchical divisive algorithm developed by Matteson and James (2014) has been utilised to detect multiple change points in a time series. The E-Divisive method has also been implemented by the authors in R, called ECP (N. A. James & Matteson, 2013). A few benefits make this method more usable compared to other change point detection approaches. For example, this method can be applied for univariate and multivariate observations (N. A. James & Matteson, 2013). Another advantage of this method and library is that it can detect multiple change point locations. Primarily, this can be done without any prior knowledge of the number of change points that appear in the time series (N. A. James & Matteson, 2013; Matteson & James, 2014). The last benefit is that this method is a statistical algorithm, and all the resulting change points are significant. This gives more objectivity to why a certain change appears. The algorithm for detecting change points involves several steps. First, the distances between all observations will be calculated to determine the similarity between each data point of the time series. In the next step, a permutation test is done to find out if there is a significant change. If a change point is available, the third step will be employed, where this algorithm will recursively apply itself to each segment until there are no more change points. In other words, it will separate this sequence according to the detected change point to find further change points in each sequence until no change point can be found. The final step involves merging the segments (divided sequences) using the k-means clustering algorithm with the aim of grouping similar statistical properties into clusters or a minimum segment size (Cabrieto et al., 2017; Sachdev, 2023).

By using this implementation in R, the outputs are the change points and the corresponding *P*-values, which establish the significance of these change points. The script for this implementation has been provided by a master's student, who also worked with toxicity scores. To put it differently, a few modifications were needed, especially for tuning the hyperparameters. The parameters to calculate the change point by *e.divisive* are: The *significance level*, which is commonly 0.05 with an *alpha* of 1, the *maximum number of random permutations* with 299, and *k* which is equal to null (N. James et al., 2023). Null means that only the statistically significant change points will be returned (N. James et al., 2023). The *minimum number of observations between change points* has been selected with

60 days, which is two years in this case. Since each considered year contains only 30 days (only November) in this thesis, and 2016 and 2017 are still "before" the design change, it was more important to investigate directly whether there are change points "after" the design change (2018). It must be noted that the minimal changes within the previous phase (2016-2017) are not relevant to answering the research question.

The change point detection has been applied to seven time-series, including the VADER sentiment compound score as well as the Perspective API score represented as *Toxicity*, *Severe Toxicity*, *Identity Attack*, *Insult*, *Profanity* and *Threat*.

3.6 Phase 3: Data Analysis

After introducing the data pre-processing, NLP methods and time series approaches, this section describes how these methods have been applied to the data to answer the research question. In this section, four different types of analysis will be described, namely a general descriptive analysis, a content-based analysis, a user-based analysis and a submission-based analysis.

3.6.1 Descriptive Analysis

The first part is the descriptive analysis of r/relationships and r/relationship_advice over the years. The following measurements have been calculated to achieve this:

- Community size per year
- Total number of collected comments per year
- Total number of active and unique users per year
- Total number of submissions per year
- Average number of comments per submission per year
- Average number of collected comments per user per year
- Average comment length per year
- Average user Karma score per year

In addition, multiple time series have been plotted. The aim was to find if there are certain trends or anomalies over time. This has been done for the following descriptive metrics:

- Total number of collected comments per year
- Average comment length per year
- Average number of comments per user per year
- Average user Karma score per year
- Total number of submissions per year
- Average number of collected comments per submission per year

This descriptive analysis aims to provide an overview of both subreddits and tries to investigate the impacts of the design change in the descriptive view.

3.6.2 Content-Based Analysis

In the content-based analysis, the focus is on analysing the comments by using both sentiment and toxicity analyses. Again, descriptive analysis was also conducted by calculating the average VADER compound score, as well as the average Perspective API score for each year. Also, the total number of collected comments and the percentages that have been classified as positive, negative, or neutral, as well as the remaining Perspective API score, have been calculated. Similar to the descriptive analysis, a time series analysis for the daily average scores for the compound score and the Perspective API attributes score have been plotted for both subreddits. Furthermore, a change point detection has been applied to each time series and has also been plotted for the corresponding plot. This analysis aims to discover across the years if a change in the degree of toxicity or sentiment takes place "after" the design and how it has been developed further.

3.6.3 User-Based Analysis

The third analysis relates to the individual users themselves. The aim was to discover how the commenting behaviour based on the toxicity of individual permanently active users has changed over time "before" and "after" the design change. This was analysed for both subreddits individually and then compared. An important criterion in the analysis was that, for example, user A in subreddit A is permanently active, meaning the user has posted at least one comment every year between 2016-2022. If the user A wrote more than one comment, the average score of the toxicity value was calculated. Since there were about 249 unique and permanently active users in r/relationships and 45 in r/relationship_advice over all the years, it was decided not to inspect each user individually but to collapse everything into one number per year. To put it differently, the average toxicity values of all permanently active users have been calculated for each year on a user basis, which is considered for r/relationships and r/relationship_advice, respectively. The results have been plotted in a line plot. This analysis aims to investigate how the design change affects the commenting behaviour of the active users when the user community stays the same over the years (closed community). In addition, the number of unique users classified as toxic has also been calculated and displayed in a number and percentage. If a user comments more than one comment, the mean toxicity score has been calculated. This analysis aims to determine how many unique users are still classified as toxic over the years "after" the design change.

3.6.4 Submission-Based Analysis

This approach aims to discover how the commenting behaviour around toxicity changes when the same user is posting the same submission on both subreddits. In this case, the focus is on the comments of submissions for the purpose of finding out how the design change has impacted both communities. In general, the aim was to find out which user community is more toxic by comparing

the comments in the same submitted submission but in different communities. The important criteria are the following:

- The submission was posted in the defined time range (each November between 2016-2022)
- The submission needs to have at least one comment
- The user must have posted this submission in both communities: r/relationships and r/relationship_advice

After this has been done for each submission, the toxicity scores for all comments were calculated and then averaged. To compare the scores of the subreddits, a counter was set for each subreddit. For example, if subreddit A were higher than subreddit B, it would increase its count. Otherwise, the counter for B would be increased. This has been done yearwise. Additionally, the overall mean score for all comments in all submissions for each year has been calculated for both subreddits. Both results have been illustrated as bar plots. In the end, 483 same submissions have been posted in both subreddits.

3.7 Implementation Details

The scripts for the data collection, pre-processing and analysis have been implemented in Python 3.9.6[1], while the change point detection has been done in R 4.3.2[2]. For the calculation of the clutter score, the libraries visual clutter[3] and Pillow Image[4] have been applied. Moreover, for analysing and pre-processing Pandas [5], NumPy [6], Seaborn[7], Matplotlib[8], Emoji[9] and datetime[10] have been used. For the sentiment analysis, vaderSentiment[11] has been applied and for the change point detection the R libraries ecp[12] have been used. The scripts and plots from this thesis can be found under: https://github.com/anthonyhami/Master_Thesis_BWT_Disorder_Toxicity.git

[1] https://www.python.org
[2] https://www.r-project.org
[3] https://pypi.org/project/visual-clutter/
[4] https://python-pillow.org
[5] https://pandas.pydata.org
[6] https://numpy.org
[7] https://seaborn.pydata.org
[8] https://matplotlib.org
[9] https://pypi.org/project/emoji/
[10] https://docs.python.org/3/library/datetime.html
[11] https://pypi.org/project/vaderSentiment/
[12] https://cran.r-project.org/web/packages/ecp/index.html

4 Results

4.1 Descriptive Results

This section shows the outcome of the descriptive analysis of the subreddits r/relationships and r/relationship_advice. These results are separated into two sections. Section 4.1.1 presents a general overview of both datasets, where the results are shown together in Table 5 and 6. Section 4.1.2 displays the descriptive results in multiple time series plot views, which can be seen in Figures 3 to 8.

4.1.1 General Descriptive Results

In Table 5, the first column defines the two subreddits, whereas the corresponding year is reported in the second column. The year shown in this table represents the month of November for the mentioned year. The third column displays the number of subscribers for each year until November since this month was chosen to be analysed. The subscribers who joined in December are not considered. These numbers are provided by Subreddit-stats (2023). The last three columns show the total number of collected comments, active users in that time range and the total number of commented submissions.

It can be observed that for both subreddits, the number of new users has increased over the years. For r/relationships, especially in the year 2018, when the design change was applied, the biggest increase in the number of subscriptions appeared. This was around two times higher than in 2017. For r/relationship_advice, the increase in the community size over the years was more drastic, with at least one million subscribers, while for r/relationships, there was a slight increase in the number of users since 2020. It is also important to mention that since 2020, r/relationship_advice had a higher community size than r/relationships. It can also be seen that "before" the design change (2016 and 2017), in both subreddits, there was an increase in active users, who commented more. For instance, in 2016, the number of users in the r/relationships community was 564k, with a total number of 225k and 32k for comments and active users, while the number of users for the same subreddit in 2017 was 884k with 253k and 36k for comments and active users, respectively. A similar trend can be detected in r/relationship_advice for the community size, total number of comments, number of active users and the total number of posted submissions.

Nevertheless, in the time range "after" the design change (2018), two different trends can be observed in the number of active users who post comments or submissions. In the case of r/relationships, since 2018 and 2019, individual users have been way less active (from 38k to 34k) and have also commented less (from 207k to 162k), although the number of new subscribers was growing for each year (from 1.8m to 2.7m). "After" the design change (2018), the number of comments (162k, 110k, 122k and 118k) and submissions (11k, 8k, 6.7k and 6.8k) never reached the corresponding numbers (200k for comments and 13k for submissions), that were "before" the design change, even with lower community size. Further investigation showed that by observing the number of active users "before" the change, 4%-5% of the users have actively participated

in the community. At the same time "after" the change, the number decreased in 2018 to 2% and was lower than 1% since 2020. Especially for the total number of collected comments and submissions, the values have dropped by about 50% by comparing 2016 with 2022. This also shows that over the years following the change, fewer active users have commented or posted, even if the community size still increased. On the contrary, for the r/relationship_advice, an opposite trend can be detected. More specifically, since 2018, more users have become active, and the number of comments and submissions has been increasing yearly dramatically. "Before" the design change, the number of active users was approximately 3%-5%, while during the year in which the design change happened (2018), 6.24% of the community was active on this subreddit. This percentage dropped from 3% to 1.2% "after" the design change, which is still higher compared to r/relationships of the same year. Nevertheless, the number of active users who commented and submitted had not fallen below the values "before" the design change since 2018 and was mostly significantly above those values. For instance, the activity by the users "after" the change, by submitting and commenting, has been increased four to five times, in comparison to the time range "before" the design change. 2021 can be seen as the year with the highest number of active users, who also posted the greatest number of comments and submissions. While in 2022, there was a decrease for all three features, these numbers are still higher than every following year since 2016. It is also interesting to mention that since the design change in 2018, especially in 2019, the community size and activity were consistently higher compared to r/relationships. However, "before" the change, it was the other way around with r/relationship_advice.

Subreddit	Year	Community Size	Total Number of Comments	Total Number of Active Users	Total Number of Submissions
r/relationships	2016	564,688	225,213	32,772	13,150
	2017	884,865	253,785	36,465	13,134
	2018	1,829,523	207,981	38,729	13,209
	2019	2,702,138	162,369	34,332	11,006
	2020	3,003,544	110,323	26,916	8,012
	2021	3,145,272	122,184	27,615	6,741
	2022	3,287,770	118,824	25,484	6,871
r/relationship_advice	2016	138,887	42,327	7,377	5,259
	2017	341,468	61,822	11,830	8,013
	2018	973,396	244,821	60,716	18,227
	2019	2,362,748	448,110	88,362	36,740
	2020	4,410,381	474,003	86,570	38,580
	2021	5,889,542	725,735	120,446	52,456
	2022	7,681,773	566,008	94,521	41,668

Table 5: **Overview of descriptive results of subreddit pair (Part 1).**

In Table 6, the second general descriptive results are displayed, which illustrate mostly comment-related features. The first two columns of Table 6 are defined exactly as described in Table 5, while the third and fourth columns outline the average number of comments per submission and per author for each year. Finally, the last two columns show the average comment length, followed by the average Karma score for each comment.

It can be seen that, in general, the average length of comments in r/relationships is higher over the years in comparison to r/relationship_advice. Furthermore, unique users, on average, are posting more comments on r/relationship_advice in comparison to r/relationships "after" the design change. By looking more in detail, it can be observed that in r/relationships, submissions were often commented on since the design change was applied, which decreased over the years. From Table

6, it can be observed that during the following two years (2019-2020), "after" the design change, the average number of comments per submission decreased. Nevertheless, since 2021, this number has become close to the number "before" the change (2016-2017). Also, it could be detected that unique active users, on average, have been writing fewer comments since the change was applied. For example, since 2019, the average number of comments per user was around four comments, while "before" the change it was around six comments. The comment length has fluctuated yearly "after" the design change. For instance, the comment length in 2018 and 2020 was higher than "before" the change. Also, the average Karma score decreased dramatically since the design change. In 2017, the average score was 19, while "after" the change it decreased until it went to 12 for the year 2022.

With regards to r/relationship_advice, a different trend can be observed. "After" the design change, the average comment score increased by a factor of 3 and no longer fell to the values it was "before" the design change. On average, unique users also posted shorter comments "after" the design change. To be precise, the comment size went from 67 to 58 words and no longer exceeded 60 words. Also, the average number of comments per submission almost doubled "after" the design change in comparison to the years "before" the change. More specifically, in 2016-2017, on average, five comments were written per submission, which doubled to 10 comments in 2018. It can be further observed that in 2018, the lowest number of comments per active user (4.03) was obtained. This number fluctuated over the years, which implies that no trend can be recognised.

Subreddit	Year	Average Number of Comments per Submission	Average Number of Comments per User	Average Comment Length	Average Karma Score
r/relationships	2016	12.28	6.87	75.93	18.61
	2017	13.79	6.96	76.12	19.51
	2018	11.77	5.37	77.07	16.41
	2019	11.17	4.73	73.41	13.16
	2020	10.18	4.10	77.19	14.03
	2021	13.46	4.42	72.96	13.96
	2022	12.56	4.66	70.56	12.21
r/relationship_advice	2016	5.87	5.74	69.41	3.25
	2017	5.94	5.23	67.94	4.22
	2018	10.41	4.03	58.15	12.41
	2019	9.85	5.07	58.22	9.78
	2020	10.02	5.48	59.68	10.74
	2021	11.38	6.03	56.78	12.12
	2022	11.27	5.99	56.89	13.03

Table 6: **Overview of descriptive results of subreddit pair (Part 2).**

4.1.2 Time Series Descriptive Results

This section visually presents the previously shown results for time series, and not only the average numbers. These time series plots are created and calculated for each day, in which the red colour represents the subreddit r/relationships, whereas the green colour represents r/relationship_advice. All the following six plots are structured in the same way, where the x-axis defines the time range corresponding to the analysed years. Particularly, one year represents the data for one month (November, as discussed in Section 3.3), which is displayed as 30 values. In Figures 3 and 4, these values represent the total number of comments and the total number of submissions, respectively.

Whereas the values in Figures 5 and 6 illustrate the average number of comments per submission and per unique user, respectively. Finally, Figures 7 and 8 display the average comment length and the average Karma score, respectively.

In Figure 3, it can be seen that with regards to the subreddit r/relationships, more comments were posted "before" the design change (2018) in comparison to "after" the design change. Opposite behaviour can be observed for the r/relationships_advice. Furthermore, since 2018, a high peak for r/relationship_advice can be seen, where the total number of comments was three times higher than "before" the change. In particular, it increased from under 5k comments to nearly 15k comments in total. After that, the number of total comments has grown nearly exponentially. Nevertheless, in the middle of November 2020, the total number of comments decreased for a few days and then increased again. Additionally, the green line has always been above the red curve since the design change. However, since 2018, the total number of comments for r/relationships has slowly fallen, where it reached approximately 50% of the initial value. That is to say, by comparing 2016 with 2022, this value dropped from 10k to 5k.

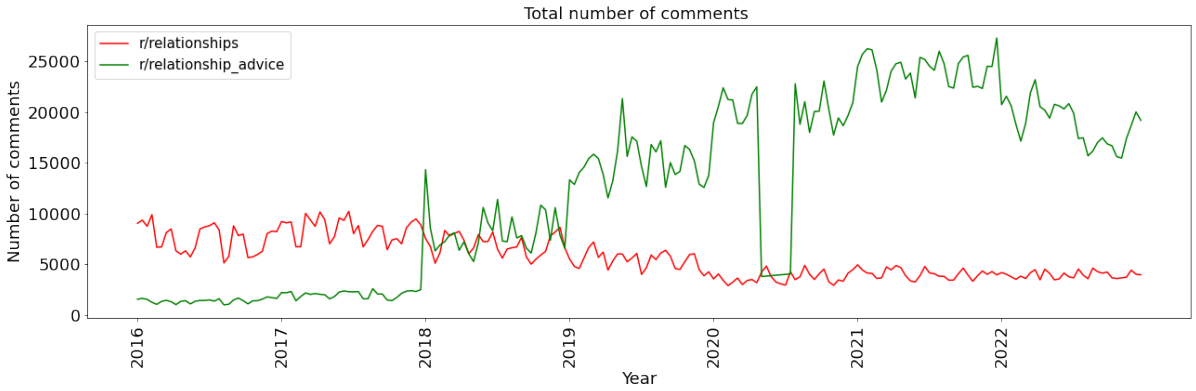

Figure 3: **Total number of collected comments per year.**

Compared to Figure 3, a similar trend can be seen in Figure 4. "Before" the change, the total number of submissions for r/relationship_advice was always lower in comparison to r/relationships. With under 500 submissions daily for r/relationship_advice, while for r/relationships, it was, on average, approximately more than 650-700 submissions daily. Nevertheless, "after" the change in 2018, a peak can be seen where the number of submissions doubled. Since then, the number of submissions that have been posted on average has also been growing higher and continuing to increase, whereas, on r/relationships, it has shrunk over the years. In 2022, it can be observed that less than 500 submissions were posted per day in r/relationships. From 2019, the number of submissions in the r/relationship_advice subreddit was always more than 1k per day. However, it can be detected that, as with the total number of comments in mid-November 2020, the total number of submissions (green line) had also fallen from more than 2k to almost 600 submissions, which rose up again. Also, it is important to mention that since the design change, more submissions have been submitted to r/relationship_advice daily. This implies that the green curve has always been above the red curve since the change, while "before" the change, it was the opposite way.

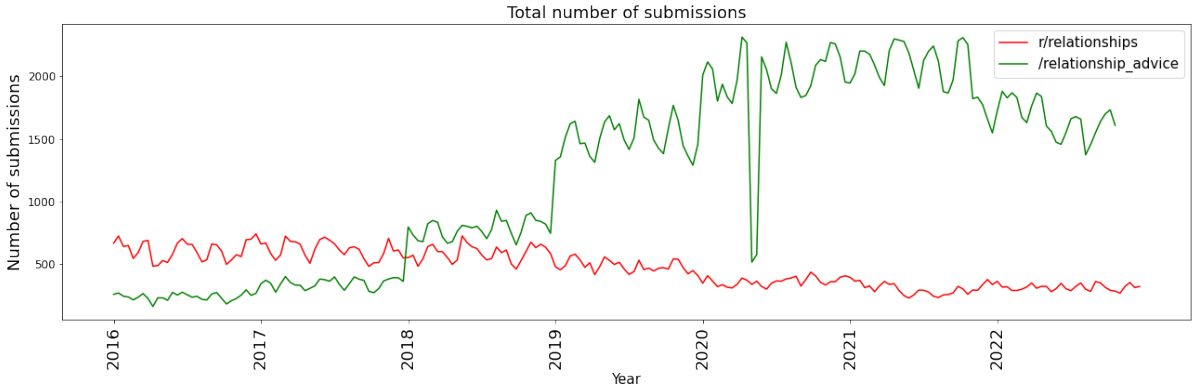

Figure 4: **Total number of submissions per year.**

Figure 5 displays the average number of comments per submission for each year (2016-2022). The y-axis defines how many comments, on average, per day have been made on one submission. It can be spotted that "before" the design change, submissions in r/relationships received almost twice as many comments per day on average than in r/relationship_advice. For instance, r/relationship_advice usually had seven comments or less, whereas r/relationships usually had between 10 and 16 daily comments in a submission. However, for r/relationship_advice, "after" the design change and at the beginning of 2018, there was a high peak, where submissions got approximately 18 comments per submission, which also decreased slightly in the same year. In both subreddits, it can be observed that the curves have fluctuated since the design change but are always close together. Nevertheless, the average number of comments on a submission for r/relationships (red curve) was always on top of the r/relationship_advice (green) curve. It is also noticeable that since the change, the r/relationship_advice curve has not fallen back to the value range it was in "before" the design change.

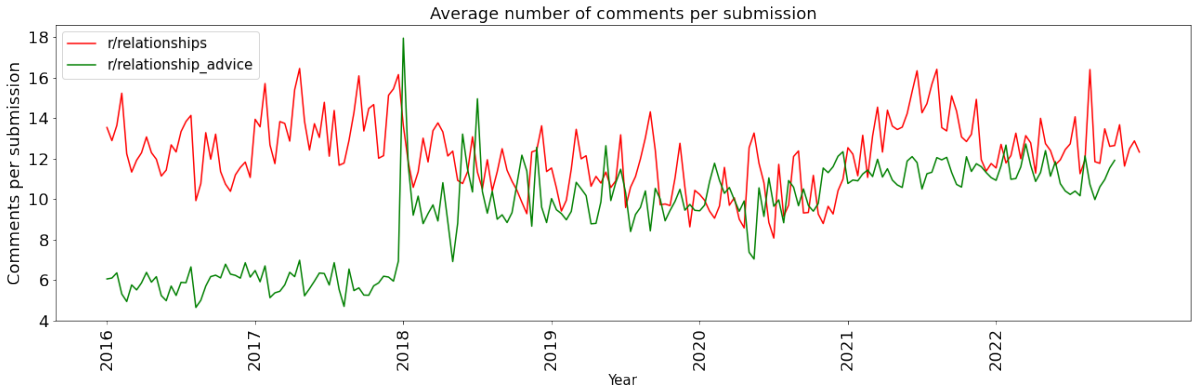

Figure 5: **Average number of comments per submission per year.**

In Figure 6, the average number of comments per user on a daily basis for 2016-2022 is presented as a time series plot. "Before" the change, individual users wrote almost the same number of comments in r/relationships as in r/relationship_advice. Both curves fell steadily and simultaneously during the time "before" the change (2016-2017) was applied. "After" the design change in 2018, both

curves dropped from almost 2.75 comments to 2 or lower. However, r/relationship_advice has seen a slight increase since 2019, and it reached the point it was "before" the change (2016-2017), while r/relationships has continued to drop. Since mid-November 2020, more unique users have written more comments in r/relationship_advice compared to r/relationships.

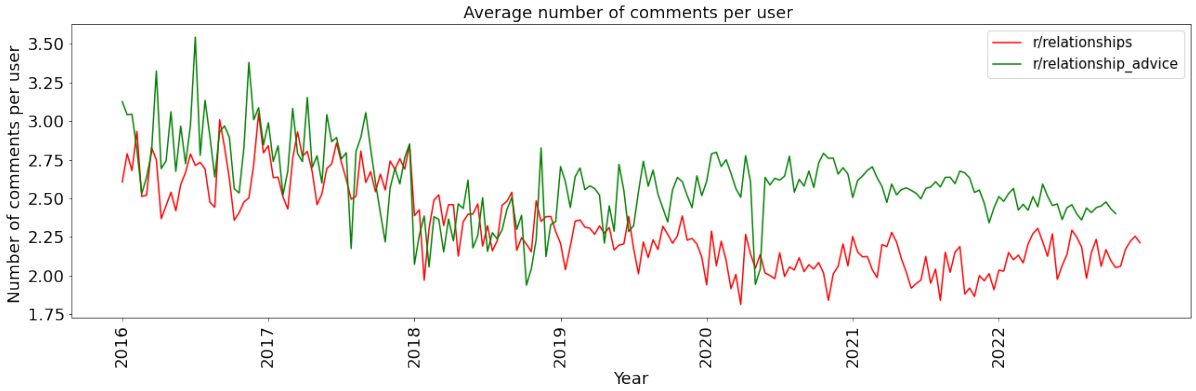

Figure 6: **Average number of comments per user per year.**

Figure 7 displays the average comment length on a daily scale. It can be seen that the average comment length in r/relationships was overall higher than the average comment length for r/relationship_advice "before" and "after" the design change. Nevertheless, at the beginning of 2017, a high peak for r/relationship_advice can be seen, where the average comment length on this day was more than 80 words long, which also decreased a few days later. To be specific, the average comment length "before" the change in r/relationships was closely between 70-80 words, while for r/relationship_advice it was 60-72 words. When the design change happened in 2018, a trend can be seen, where the average comment length dramatically reduced from 70 to nearly 56 words for r/relationship_advice. Since then, it stood between 55 and 60 words until 2022 and never reached the average comment length it was "before" the change. As for the subreddit r/relationships, nothing changed much over the years, and much fluctuation happened, where the average comment length stayed mostly between 70 and 80 words. In mid-November 2021, a high decrease can be observed (red line), where the average comment length corresponds to less than 65 words, yet, increased back to more than 70 words.

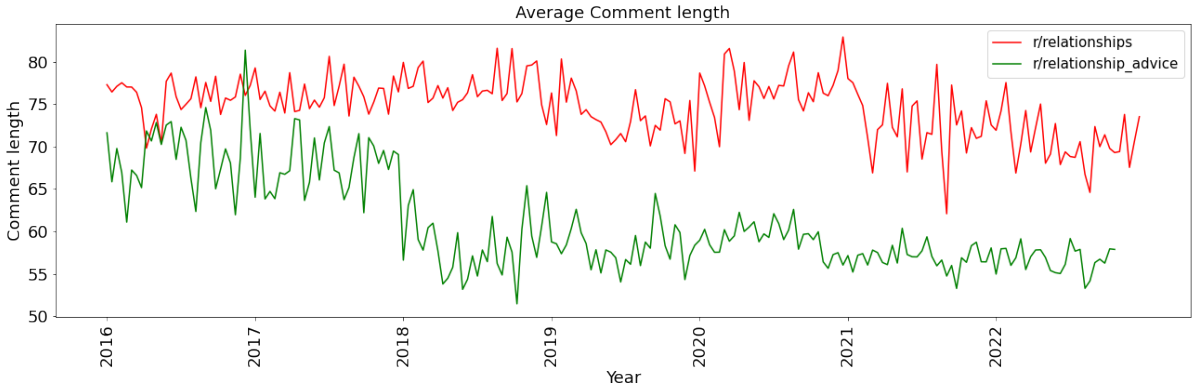

Figure 7: **Average comment length per year.**

Figure 8 shows the average voting score on a daily scale for 2016-2022. It is noticeable that the average Karma score "before" the change was mostly under five votes for r/relationship_advice, while for r/relationships it was nearly four times higher with approximately over 15-20. "After" the change, the average score for r/relationships slowly decreased over the years until it reached the ten-vote average boundary in 2022. On the contrary, for r/relationship_advice, there was an increase at the beginning of 2018, with an average vote growing up to 3 times higher than for the r/relationships subreddit. Especially near the end of 2018, a high peak with more than 40 votes can be seen. This later decreased and stayed between 10 and 15 until 2022. Since 2020, the average score for r/relationships decreased, while r/relationship_advice slightly increased.

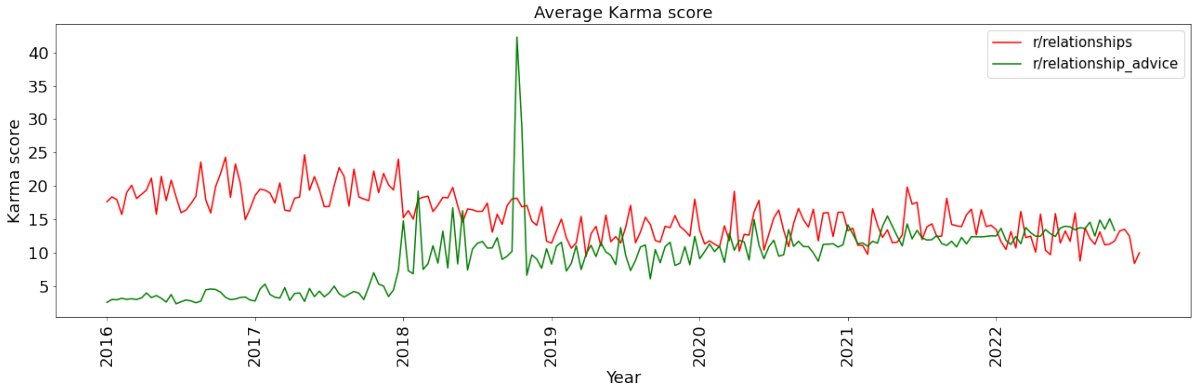

Figure 8: **Average Karma score per year.**

4.2 Content-Based

This section represents the results of the content-based analysis, which are divided into two parts. Section 4.2.1 illustrates the sentiment results where Table 7 shows the descriptive results, and Figure 9 shows the time series plot for the sentiment analysis, in which the change point detection has been applied. Section 4.2.2 demonstrates the toxicity analysis, where Tables 8 and 9 show the descriptive analysis for toxicity and Figures 10-12 present the time series plots for the toxicity attributes, in which also the change point detection has been applied.

4.2.1 Sentiment

In Table 7, the descriptive sentiment analysis results are displayed, where the first column defines the subreddit and the second column shows the results for the corresponding years. In columns 3-5, the number of positive, neutral and negative comments are shown. Additionally, the percentage of each of these categories is shown in the parentheses. For each row, the percentage should add up to 100. The last column represents the average VADER compound score over those years. "Before" the design change, the number of positive comments was similar for both subreddits, while in 2017 r/relationship_advice had a slightly higher percentage of positive comments. This can also be noted by the compound score of 0.20 for r/relationship_advice and 0.17 for r/relationships. Regarding the number of negative comments, it can be observed that r/relationships had also slightly more negative comments with 33%, while r/relationship_advice had 31%-32%. For r/relationship_advice, more

comments were classified as neutral for all years in comparison to r/relationships. However, "after" the design change, it can be seen that for the subreddit r/relationships, the number of positive comments has increased and never dropped back to the values "before" the change. In other words, it ranged between 57.50%-59% ("after"), which is an increase of 3% ("before"). Also, a similar trend can be seen by investigating the negative comments, which decreased from 33.81% ("before") to values between 31%-32% ("after"). The neutral comments fluctuated over the years, yet in 2022, the highest number was reached, with a percentage of 10.25% for the neutral comments. This trend can be summarised in the average number of compound scores as well, where in 2018, it was 0.20, which is higher than the years "before" the change. "After" the design change, this value fluctuated over the years between 0.19 and 0.22. This implies that the comments were more positive than "before" the design change. With regards to r/relationship_advice, an opposite trend can be seen, where the positive comments have decreased from 57% in 2017 to 53% in 2018, whereas the number of negative comments increased from 31.01% to 33.00%. This trend can also be seen in the average VADER sentiment score, where the score has been decreased by 0.20 to 0.15. To put it differently, "after" the change, more negative comments were written while positive comments were reduced. On the contrary, the number of neutral comments increased in 2018 by 2.5% from 2017. It is important to mention that "before" the change, the comments in the subreddit r/relationship_advice were nearly more positive and neutral in comparison to r/relationships. However, "after" the change, the commenting behaviour in r/relationships was way more positive, neutral, and less negative than for r/relationship_advice. Generally, the years in which the comments were mostly positive and less negative for r/relationships and r/relationships_advice were 2020 and 2022, respectively.

Subreddit	Year	Number of Positive Comments	Number of Neutral Comments	Number of Negative Comments	Average VADER Sentiment Compound Score
r/relationships	2016	128,298(56.97%)	20,785(9.23%)	76,130(33.80%)	0.1755
	2017	144,692(57.01%)	23,277(9.17%)	85,816(33.81%)	0.1754
	2018	122,128(58.72%)	18,693(8.99%)	67,160(32.29%)	0.2034
	2019	194,489(58.19%)	15,464(9.52%)	52,416(32.28%)	0.1980
	2020	65,984(59.81%)	99,320(9.00%)	34,407(31.19%)	0.2246
	2021	71,891(58.84%)	11,971(9.80%)	38,322(31.36%)	0.2121
	2022	68,324(57.50%)	12,180(10.25%)	38,320(32.25%)	0.1943
r/relationship_advice	2016	23,725(56.05%)	4,747(11.22%)	13,855(32.73%)	0.1753
	2017	35,708(57.76%)	6,933(11.21%)	19,181(31.03%)	0.2029
	2018	130,485(53.30%)	33,545(13.70%)	80,791(33.00%)	0.1514
	2019	243,741(54.39%)	57,759(12.89%)	146,610(32.72%)	0.1600
	2020	259,234(54.69%)	59,632(12.58%)	155,137(32.73%)	0.1649
	2021	386,471(53.25%)	96,468(13.29%)	242,796(33.46%)	0.1480
	2022	297,735(52.60%)	75,846(13.40%)	192,427(34.00%)	0.1386

Table 7: **Overview of descriptive sentiment analysis results of subreddit pair.**

Figure 9 presents the average compound score, calculated daily for both subreddits, r/relationships being in red and r/relationship_advice being in green. The change point detection has been applied, and the corresponding detected points can be seen in vertical lines on the figure in blue and yellow for r/relationships and r/relationship_advice, respectively. Up to 2018, it can be detected that both curves are almost parallel, as they denote similar trends, yet, r/relationship_advice was at the beginning above r/relationships. In general, "before" the design change r/relationships (red), the

curve was more constant, while r/relationship_advice (green) grew slightly higher. Nevertheless, "after" the design change, r/relationships average compound score (red) increased over the years and was higher than the average compound score of r/relationship_advice (green), which itself decreased more since the design change

As a result of the change point detection, only one change point was identified for r/relationships (blue) and two change points for r/relationship_advice (yellow). For both subreddits, the change points (the first yellow and the second blue) were identified in 2018, when the design change occurred. For r/relationships, the change point was on 13.11.2018, with a *P*-value of 0.0033, whereas for r/relationship_advice, it was on 01.11.2018 and 01.11.2020 with the *P*-values of 0.0033 and 0.0133, respectively. By focusing on the change points, it can be observed that for r/relationships, the curve has increased higher since then. However, for r/relationship_advice, the first change point shows a decrease indicating that the comments were more negative since the design change. The average compound scores between the first and second change points for r/relationship_advice stay constant apart from some fluctuation. After the second change point, another decrease occurred. All in all, it can be seen that "after" the design change, r/relationships went more positive, while r/relationship_advice got more negative on average over the years 2018-2022.

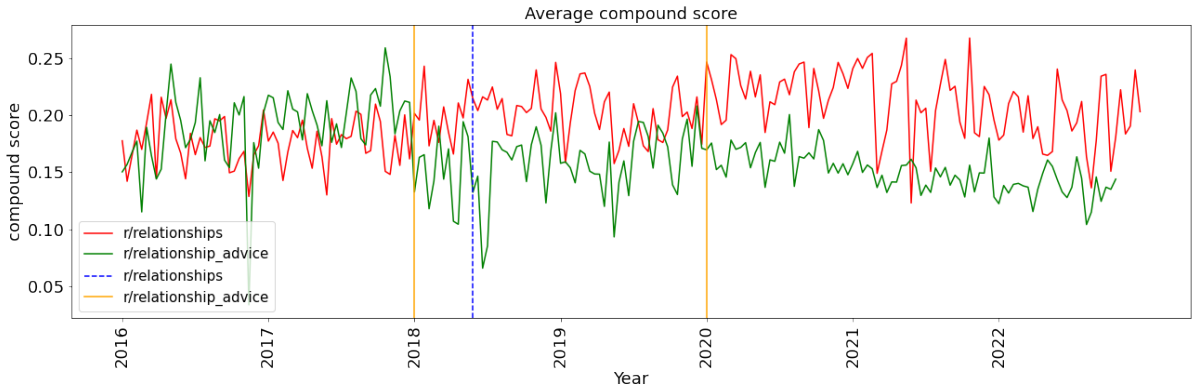

Figure 9: **Average compound score per year with change points.**

4.2.2 Toxicity

This section presents the results of the content-based toxicity analysis. Based on the toxicity analysis, it could be found that for the categories *Severe Toxicity*, *Identity Attack* and *Threat*, few comments were recognised or that the score was usually very low for these categories. Therefore, it was decided to only focus on toxicity analysis in the categories of *Toxicity*, *Profanity* and *Insult*. The results for the other Perspective API categories can be seen in Appendix E.

Table 8 displays the yearly average scores for the attributes *Toxicity*, *Profanity* and *Insult*. The first two columns have a similar structure as defined in previous tables, while the last three columns show the average score for the Perspective API attributes. The remaining three attributes can be found in Appendix E in Table 13. It can be observed that "before" the change in 2016, r/relationship_advice had, on average, higher *Toxicity* (0.22) and *Profanity* score (0.16) than r/relationships (0.21 and

0.14, respectively), while the *Insult* score (0.12) stayed the same. On the contrary, in 2017, for both subreddits, the scores decreased or stayed constant, but r/relationship_advice had in that year a higher *Profanity* score (0.15) compared to all years for r/relationships. For r/relationships, all average Perspective API scores for 2016-2017 decreased or stayed constant in the year where the design change happened (2018). In contrast, the average scores during the design change (2018) for r/relationship_advice has been increased by 0.02-0.03. These scores (0.23, 0.17 and 0.13) in 2018 are the highest for this subreddit for all years. It can also be detected that "after" the design change, r/relationship_advice was for all toxicity attributes more toxic in comparison to r/relationships. On the one hand, the year where r/relationships had the lowest score overall was 2022 for each category, while the highest score was 2016, which was "before" the change. On the other hand, 2018 was for r/relationship_advice, the year which had the highest score for all Perspective API scores. All in all, since 2019, the score for both subreddits has stayed constant or changed slightly by a score of 0.01.

Subreddit	Year	Average Toxicity	Average Profanity	Average Insult
r/relationships	2016	0.21	0.14	0.12
	2017	0.20	0.14	0.12
	2018	0.19	0.13	0.11
	2019	0.19	0.13	0.11
	2020	0.18	0.12	0.11
	2021	0.19	0.12	0.11
	2022	0.19	0.12	0.11
r/relationship_advice	2016	0.22	0.16	0.12
	2017	0.20	0.15	0.11
	2018	0.23	0.17	0.13
	2019	0.22	0.15	0.12
	2020	0.21	0.15	0.12
	2021	0.22	0.15	0.12
	2022	0.22	0.14	0.12

Table 8: **Overview of average Perspective API score per year of subreddit pair.**

Figure 10, 11 and 12 illustrate the time series results for the attributes *Toxicity*, *Profanity* and *Insult* for both subreddits, respectively. These Figures have a similar structure as described in the section 4.2.1 the results for content-based sentiment analyses.

Figure 10 shows the average *Toxicity* score per year on a daily basis for both subreddits. "Before" the change (2016-2017), both curves are close together, whereas r/relationship_advice is, on average, slightly higher, meaning more toxic. The average values for r/relationships are between 0.20 and 0.22, whereas for r/relationship_advice, the values are between 0.18-0.23. Nevertheless, the *Toxicity* scores for r/relationship_advice (2016-2017) are mostly decreasing, while for r/relationships they stayed mostly constant. "After" the design change, the average score for r/relationship_advice increased and then stayed constant, while for r/relationships it decreased with some fluctuations. This can be seen as the change point detection could observe two change points for r/relationship_advice (2018 and 2020), while for r/relationships there is only one change point (mid-November 2018). For r/relationships, the change point was found on 13.11.2018, with

a *P*-value of 0.003. For r/relationship_advice, the first change point was detected on 01.11.2018, with a *P*-value of 0.0033, and the second change point was detected on 01.11.2020, with a *P*-value of 0.0133. Since then, the average *Toxicity* score has not changed anymore and stayed consistent, but is still higher than "before" the change. It is also important to mention that since the design change, the curve of r/relationship_advice has always majored the curve of r/relationships, which indicates that the average *Toxicity* score was higher in comparison to r/relationships.

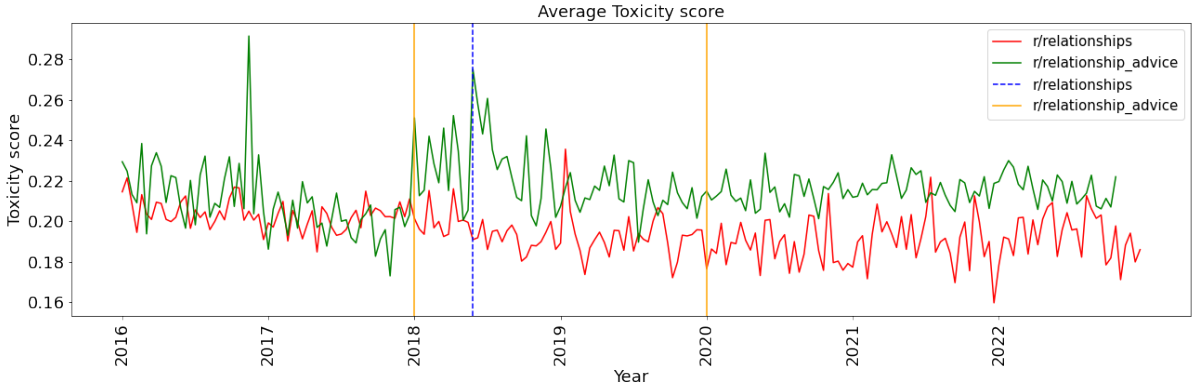

Figure 10: **Average *Toxicity* score per year with change points.**

A similar trend can be observed for *Profanity* and *Insult* in Figures 11 and 12, respectively. Yet, the detected change points are different, especially for *Profanity*, while for r/relationships it is on 16.11.2018 with a *P*-value of 0.0066, and for r/relationship_advice it is on 17.11.2019 with a *P*-value of 0.00333.

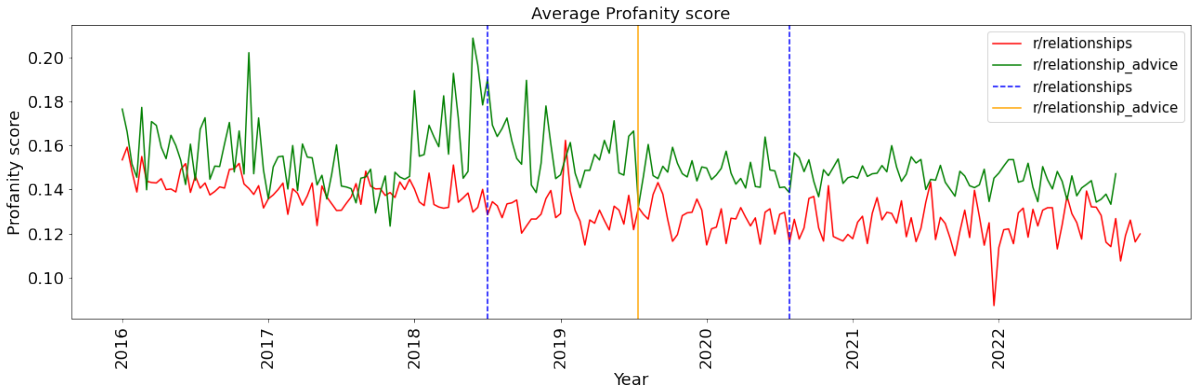

Figure 11: **Average *Profanity* score per year with change points.**

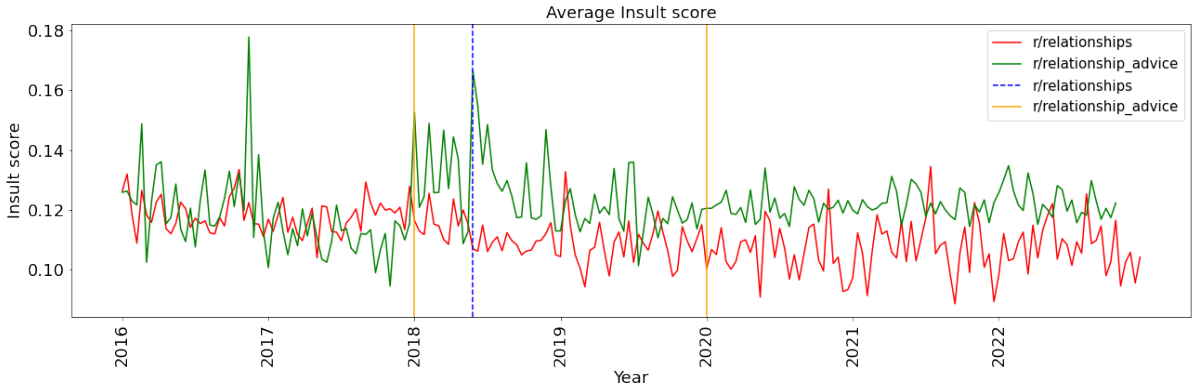

Figure 12: **Average *Insult* score per year with change points.**

Table 9 presents the number of classified comments of the attributes *Toxicity*, *Profanity* and *Insult*, as well as their corresponding percentages. The remaining attributes can be found in the Appendix E in Table 14. "Before" the design change (2016-2017), the percentage number of comments that were classified as toxic for r/relationships (4.21% and 3.78%) were lower than the percentages for the r/relationship_advice (6.38% and 4.96%). Within the subreddits, the percentage scores decreased from 2016 to 2017. Apart from 2021, the percentage number of comments classified according to the different Perspective API scores for r/relationships has continuously decreased over the years. This indicates that fewer *Toxicity*, *Profanity* and *Insult* comments were written in this subreddit.

However, the percentage numbers for the r/relationship_advice "after" the design change increased, which shows an opposite trend compared to r/relationships. 2018 reported the highest number of those percentages. Since 2019, the percentage number of comments decreased until 2022. Except for 2018, a decrease in toxic comments "after" the change was observed. For example, in 2016, the percentage of *Profanity* comments was 5.86%, while in 2022, this number went to nearly 4%. Those numbers are still lower than the percentage of 6.82% in 2018.

Subreddit	Year	Number of Toxic Comments	Number of Profanity Comments	Number of Insult Comments
r/relationships	2016	9,382(4.21%)	9,384(4.20%)	1,767(0.79%)
	2017	9,341(3.78%)	9,127(3.60%)	1,942(0.77%)
	2018	7,246(3.58%)	7,013(3.37%)	1,420(0.68%)
	2019	5,584(3.43%)	5,373(3.30%)	1,141(0.70%)
	2020	3,472(3.14%)	3,246(2.94%)	696(0.63%)
	2021	4,090(3.34%)	3,648(2.98%)	864(0.70%)
	2022	3,789(3.19%)	3,279(2.76%)	775(0.65%)
r/relationship_advice	2016	2,657(6.38%)	2,481(5.86%)	458(1.08%)
	2017	3,067(4.96%)	2,995(4.86%)	558(0.90%)
	2018	18,098(7.40%)	16,694(6.82%)	4,046(1.65%)
	2019	25,991(5.80%)	23,836(5.32%)	4,869(1.09%)
	2020	26,041(5.49%)	22,267(4.70%)	5,925(1.25%)
	2021	38,853(5.35%)	33,079(4.55%)	8,014(1.10%)
	2022	27,914(4.93%)	22,861(4.04%)	6,134(1.08%)

Table 9: **Overview of number of comments classified as *Toxic*, *Profanity* and *Insult*.**

4.3 User-Based

This section includes the results of the user-based analysis, which are also based on toxicity and are divided into two ways. Table 10 shows the results of how many unique users were classified as toxic over the years, while Figures 13 to 15 display the results of all permanent active users that commented between 2016 and 2022 by averaging the Perspective API values into one value for each year. Both analyses have been done for the subreddit r/relationships and r/relationships_advice separately. Similar to the previous analysis, the focus is on the Perspective attributes of *Toxicity*, *Profanity* and *Insult*. The remaining results for the other attributes can be found in Appendix F in Table 15 and in Figures 34 to 36.

In Table 10, it can be observed that "before" the design change (2016 and 2017), a slight increase in the number of unique users that have been classified as *Toxicity*, *Profanity*, and *Insult* for both subreddits could be found. However, by looking at the percentages of r/relationships_advice for *Profanity* (2.68% to 2.23%) and *Toxicity* (2.72% to 2.20%), it slightly decreased between those years. Also, it can be detected that for r/relationships, more users could be identified as *Toxicity*, *Profanity*, or *Insult*, in comparison to r/relationships_advice.

During the design change (2018) for r/relationships, the percentage of users classified as toxic in the different categories was lower than the years before. Since the design change (2018) for r/relationship_advice, the percentage of unique users classified as one of the categories decreased each year. It is to be noted that there is a 4% increase in toxic unique users between 2017 and 2018 (2.20% to 6.20%). However, after 2018, the percentage share decreased for all categories over the year. In 2022, only 2.54% of active users were classified as toxic, which is still slightly more than in 2017 "before" the change. In the end, even "after" the change, r/relationships still had fewer toxic users compared to r/relationship_advice over the years.

Subreddit	Year	Number of Unique Toxic Users	Number of Unique Profanity Users	Number of Unique Insult Users
r/relationships	2016	623(1.90%)	634(1.93%)	104(0.32%)
	2017	729(2.00%)	724(1.98%)	145(0.40%)
	2018	681(1.75%)	685(1.76%)	115(0.29%)
	2019	761(2.21%)	694(2.02%)	133(0.38%)
	2020	530(1.96%)	487(1.80%)	89(0.33%)
	2021	572(2.07%)	489(1.77%)	119(0.43%)
	2022	507(1.98%)	431(1.69%)	104(0.41%)
r/relationship_advice	2016	201(2.72%)	198(2.68%)	32(0.43%)
	2017	260(2.20%)	264(2.23%)	54(0.46%)
	2018	3,753(6.20%)	3,405(5.61%)	853(1.40%)
	2019	3,563(4.03%)	3,218(3.64%)	701(0.79%)
	2020	2,687(3.10%)	2,177(2.51%)	605(0.69%)
	2021	3,699(3.10%)	3,012(2.50%)	777(0.64%)
	2022	2,405(2.54%)	1,913(2.02%)	528(0.55%)

Table 10: **Overview of number of unique users that classified as *Toxic*, *Profanity* and *Insult*.**

In Figures 13 to 15, the lines represent the mean of the average Perspective scores and the corresponding standard deviation (the area above and below the lines). The average Perspective scores are computed for all users (in the community), who commented at least one comment through the years 2016-2022. In total, 29,743 comments have been made by the same user community in r/relationships, while in r/relationship_advice it was 46,462 comments. There were about 249 unique and permanent active users in r/relationships and 45 in r/relationship_advice.

In Figure 13, the user-based results of the closed community on *Toxicity* can be seen. It can be observed that until 2020, the users in r/relationships had, on average, a higher score in comparison to r/relationship_advice, except for the year 2021, where the r/relationships score was lower than for r/relationship_advice. Overall, these scores (2016-2021) are decreasing during those years. For r/relationship_advice, there is a decrease from 2016-2017 (0.2 to 0.16), to which an overall increase follows. Those numbers, however, do not cross the score in 2016 (0.2). "After" the change in 2018, there was a large increase from 0.16 (2017) to 0.18 (2018). The standard deviation for the r/relationship_advice is higher than for r/relationships, which could indicate that the user's behaviours are different.

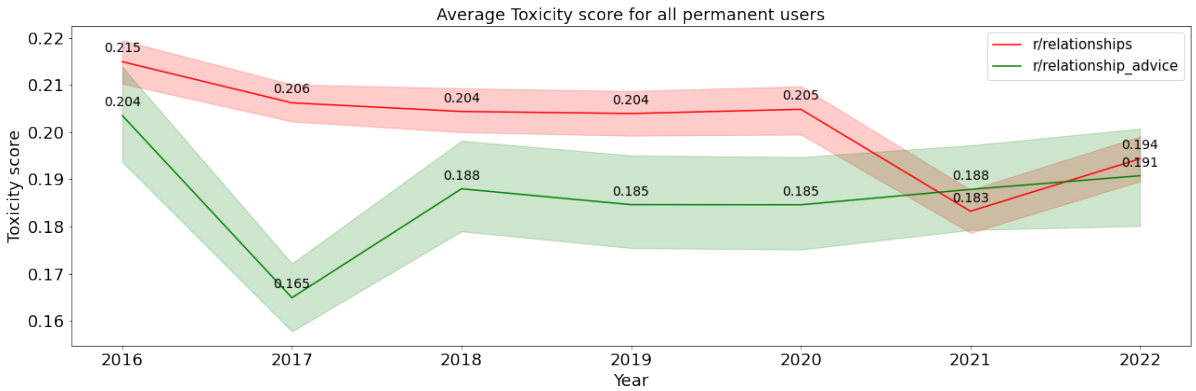

Figure 13: **Average *Toxicity* score for all permanent active users.**

The same structure can be seen in Figures 14 and 15 for *Profanity* and *Insult*, respectively. One difference is in Figure 15, where in 2021 the average *Insult* score in r/relationships was higher than in r/relationship_advice. Another difference is that in Figure 14, the average *Profanity* score for r/relationship_advice was higher than r/relationships. Generally, the *Toxicity* score range is higher than the *Profanity* and *Insult* ranges.

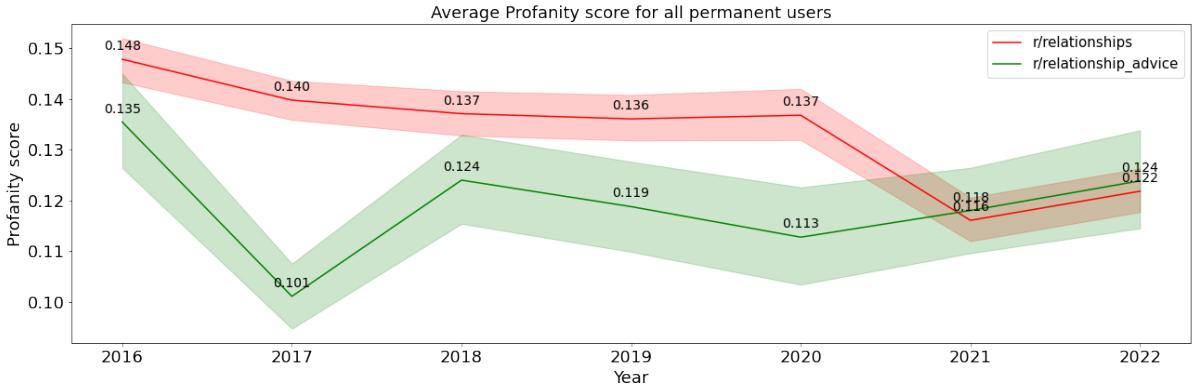

Figure 14: **Average *Profanity* score for all permanent active users.**

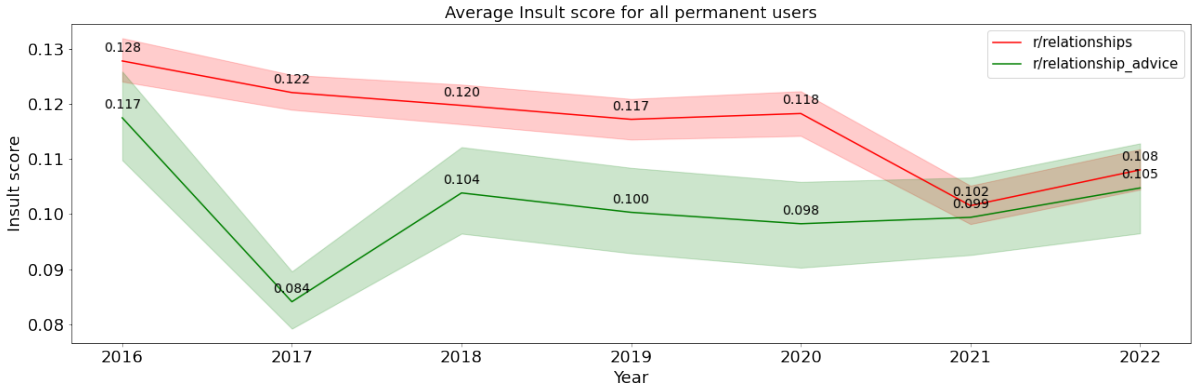

Figure 15: **Average *Insult* score for all permanent active users.**

To summarise, users in the closed community of r/relationships became less toxic or stayed stable in comparison to "before" the change. For r/relationship_advice, the toxicity score in 2018 is similar to the score in 2022.

4.4 Submission-Based

In this section, the submission-based results are presented. The results are shown in two figures, where Figure 16 displays the *Toxicity* count and Figure 17 displays the average *Toxicity* score of all comments from each submission. The *Toxicity* count is computed as follows: The average *Toxicity* score for each comment in a submission is computed for both subreddits and compared. A count will be added to the corresponding subreddit if the computed average score is higher for that subreddit (see Section 3.6.4). Since the results for *Toxicity*, *Profanity* and *Insult* look very similar, only the *Toxicity* score is shown here and the remaining plots can be found in Appendix G (Figures 16-46).

In Figure 16 and up until 2018, the counted toxic submissions for r/relationship_advice were higher than for r/relationships. Yet, since 2019, the r/relationships community has been seen to be more toxic in their submissions compared to r/relationship_advice.

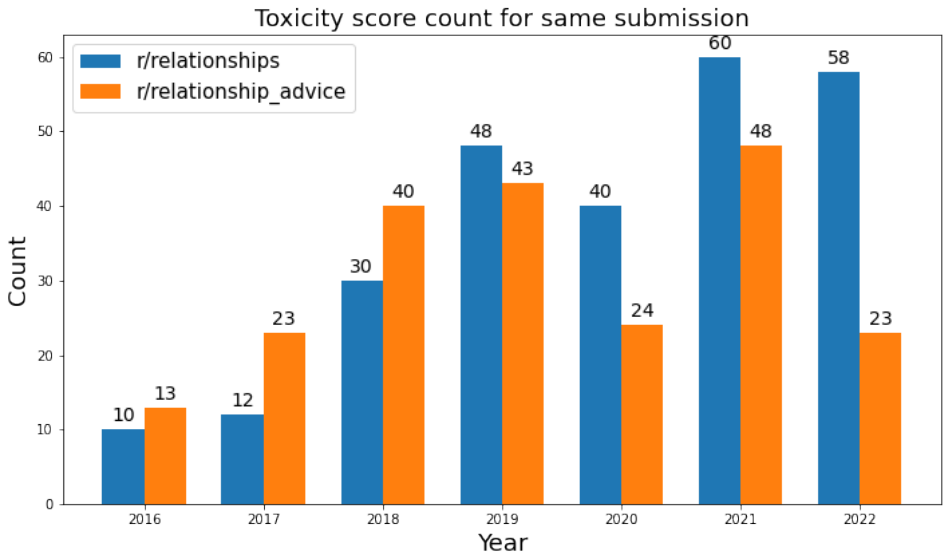

Figure 16: ***Toxicity* count for submission-based analysis.**

This trend can also be observed in Figure 17, where the average score for all comments has been calculated. However, the average score for r/relationships is larger than for r/relationship_advice from 2020 (and not 2019) to 2022.

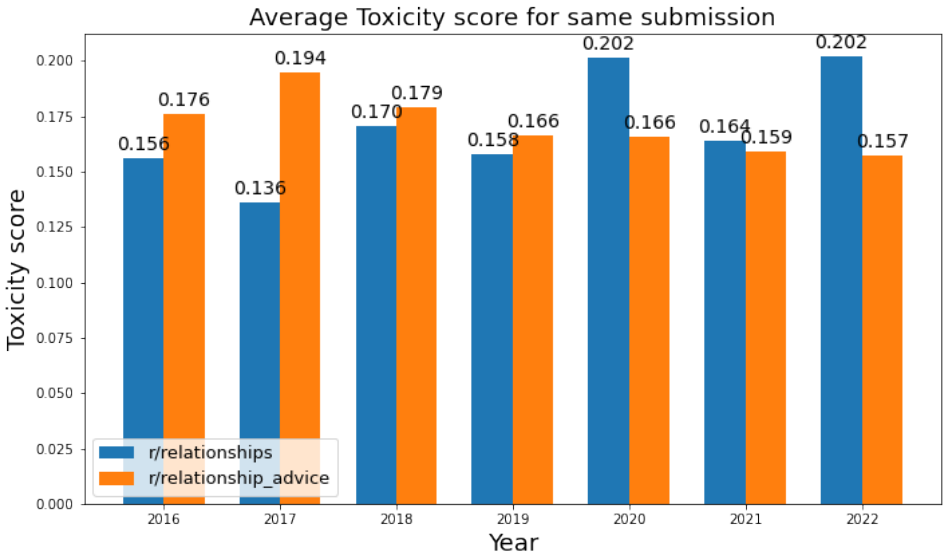

Figure 17: **Average *Toxicity* score for submission-based analysis.**

5 Discussion

5.1 Summary of the Data Analysis Results

This thesis aims to investigate the BWT in an online environment by finding out how the toxic and general commenting behaviour would change when a "badly designed" and "disordered" website is redesigned to a more structured and less disordered page. This has been realised with a 3-steps workflow (described in Section 3.1), where at the end, four different data analysis views have been employed to answer the research question and the four pre-defined hypotheses:

RQ: Does perceived disorder in an online environment increase toxicity?

H1: The toxic commenting behaviour of users in an online environment (Reddit) will improve after a design change from disordered to less disordered.

H2: The number of active and unique toxic users will decrease after a design change from disordered to less disordered.

H3: All users, who were permanently active within a defined time range, will decrease their toxic behaviour after a redesign.

H4: The subreddit community that appears to be more disordered after the redesign will reflect higher toxic behaviour in comparison to the subreddit which is less disordered.

Descriptive Analysis This analysis demonstrates that the number of new users joining both subreddits (r/relationships and r/relationship_advice) is increasing over the observed period of time (2016-2022). This is mainly observed after the design change (2018). However, the number of comments and submissions decreased "after" 2018 for r/relationships. An opposite behaviour is seen for r/relationship_advice. Additionally, the length of the comments in r/relationship_advice decreased, yet, the number of comments increased. The Karma score for r/relationships decreased, while the score increased for r/relationship_advice.

Content-Based Analysis The change point analysis (compound score "after" the design change) shows that the r/relationship_advice environment becomes more negative, whereas the r/relationships environment becomes more positive. Also, the toxicity analysis showed similar behaviour, where in r/relationships the number of toxic comments decreased, while in r/relationship_advice it increased until 2021, where it stabilised and decreased slightly between 2021-2022.

User-Based Analysis In this analysis, r/relationships "after" the redesign had fewer unique toxic users, while for r/relationship_advice, the number of toxic unique users increased until 2021, and then the number stabilised and decreased. For the permanent active user analysis (especially

"after" the change), the average toxicity decreased slightly or stayed unchanged in r/relationships, while in r/relationship_advice, the average score increased and stayed mostly stable with minimal changes. However, r/relationships had a higher average toxicity score over the years compared to r/relationship_advice.

Submission-Based Analysis The results showed that between 2018 and 2019, users from the r/relationship_advice community had more toxic reactions to the same submission than r/relationships. However, since 2020, r/relationships displayed a higher average toxicity score, where they commented more toxic to the same submission in comparison to r/relationship_advice.

5.2 Discussion and Implications of the Data Analysis Results

This section discusses the results and their implications. It is structured according to the following sequence: First, the descriptive results will be discussed. Second, the content-based analysis will be discussed, where the validity of the hypothesis *(H1)* will be examined as well. Third, the user-based analysis will be discussed, and the hypotheses *(H2)* and *(H3)* will be answered. Fourth, the submission-based analysis will be briefly discussed, and fifth, the hypothesis *(H4)* will be investigated, under consideration of all results previously discussed. Lastly, the research question is answered.

Descriptive Analysis This analysis has shown that "after" the design change, the number of new users for both subreddits has dramatically increased. This could indicate that the design change might be a factor in this increase of new subscribers. Furthermore, it can be assumed that Reddit's online environment has become more attractive to new users due to the redesign. In an interview with Huffman, the co-founder of Reddit, he mentioned that the aim was to "*transform the dystopian Craigslist-looking Reddit into a more modern and user-friendly environment*" (Pardes, 2018). In terms of BWT in the online context, it can also be argued that through the design change new users no longer see Reddit as a dangerous or unsafe environment, which is why the users want to visit and join the new designed platform. To confirm this statement, "u/Mercurycandie", a latter moderator (between 2019 and 2022) of r/relationship_advice was contacted to get more evidence of the user's behaviour "before" and "after" the design change. This moderator has been an active member on Reddit since 2012, especially on r/relationships. He was also moderating other subreddits. He confirmed that this trend (the increase of the community size) has also taken place in the other subreddits that he was moderating. In terms of the total number of comments, submissions and active users, it could be seen that users commented or posted fewer submissions on r/relationships "after" the design change period. In contrast, r/relationship_advice saw a steady increase in comments and submissions. One possible reason for this might be the rules set by the moderators and corresponding moderations for the respective subreddits. For instance, the rules for the r/relationships subreddit were stricter and took on a very restrictive model of who could post a submission and how long those submissions could stay up, whereas the rules for r/relationship_advice were much more flexible with less strict moderation. This could lead to an increase in the popularity of the r/relationship_advice, which ended up with submissions hitting

the front page far more often and thus led to a more rapid increase in comments and submissions. For new users, it was harder to find r/relationships, as fewer submissions appeared on the front page, which could result from the strict rules being implemented. Since the moderation was so strict on r/relationships, probably fewer comments were posted on this subreddit compared to r/relationship_advice, as a rule break can lead to a removal of the comment or a harder punishment like a ban of the user. This can be observed in the number of removed comments, where r/relationships had a high count of removed comments (see Table 16 in Appendix H). All comments and submissions that do not follow the rules of r/relationships will be immediately removed. Those moderation rules were very diverse and provided detailed information that even a separate FAQ page was given by r/relationships itself (Archive, 2024). For example, comments or submissions could be removed for various reasons such as insults, fake submissions, off-topic for the subreddit or if the format and style of the submission were not adhered to. Another strict rule for r/relationships was that submissions or comments that get a high number of votes are classified as fake and will be removed. This could also be the reason why the average score has decreased too. Meanwhile, on r/relationship_advice, all rules are listed and described in short sentences in the front-page sidebar. It is to be mentioned that these strict rules were already in place "before" the design change and were only slightly expanded "after" the change for both subreddits. Another argument why the design change did not have a huge impact on both communities could be an external factor outside of Reddit. It was found out that other online environments like X, TikTok or YouTube attached screenshots from submissions of r/relationship_advice and made fun of them or of the corresponding comments (Tait, 2020). This indicates that the high number of active users from r/relationship_advice arose because of the members of other social media platforms. Those external members only enter this community to watch or comment on those special submissions that they saw on the other platform as re-uploads. One assumption could be that a high number of active users could be "throwaway accounts", which are only created to post a comment or submission, and afterwards, those accounts will be deleted by the users themself. On the one side, the number of deleted users for r/relationship_advice was high, especially "after" the change (see Table 16 in Appendix H). On the other side for r/relationships, the number of deleted users was less. Nevertheless, it can be seen that at r/relationship_advice from 2021, the number of removed comments has also increased, and the number of deleted users has decreased. One reason for that trend could be an internal rule change in r/relationship_advice. According to the latter moderator, new rules were introduced in 2020. Those rules state that comments or submissions need to be marked as "throwaway accounts" if the user does not want to keep this account, and thus the corresponding submission or post will be removed. A further rule is that submissions or comments that get more than 2k votes will also be removed, as they are seen as fake or troll.

Content-Based Analysis Similarly, the result of the content-based analysis indicates that "after" the design change, r/relationship_advice gained more negative and toxic behaviour, while r/relationships had more positive and less toxic behaviour. This can also be observed in the time series plots, where for almost every time series a significant change point could be found in 2018 for each attribute (except *Profanity*). In general, new subscribers on r/relationships, who would

like to spread antisocial behaviour, have been discouraged by the strict rules. They, certainly, do not want to waste time or effort on this subreddit and, thus, try to find another victim where it is easier to be toxic. Whereas for r/relationship_advice, the opposite happened regarding toxic and negative behaviour. It could be assumed that more throwaway accounts have been created that had the aim to just spread trolling or toxic behaviour instead of discussing the content. This can also be seen by the decreased average comment length "after" the change, which could mean that less deliberation is included in this comment and more toxicity is included. However, since 2021 the toxic and negative behaviour has decreased or stayed stable, which is apparent in the time series plots of *Toxicity*, *Insult* and compound score, where a second change point was found in 2020. This could be a reason for the introduction of the new rules on r/relationship_advice, which also removes the fake posts with high votes. Above all, it was not possible to prove Messerschmidt et al. (2023), Atari et al. (2022) or Horta Ribeiro et al. (2021) statements that toxic behaviour in small communities is higher than in larger ones, as the opposite happened in these cases. To put it differently, r/relationship_advice, which had more users and a higher number of comments and submissions "after" the change, had a higher polarity than r/relationships. This statement could also be confirmed by Borah (2014) and Xia et al. (2020).

With this, it can be concluded that the hypothesis *(H1)* for r/relationships was proven to be true, whereas for r/relationship_advice it was not proven to be true and was always disproved. Nevertheless, all the reasons for this change of toxicity at r/relationships might be the moderation rules. In contrast, the design change might not have a huge impact as expected, as explained in the previous paragraph of descriptive analysis. Thus, *H1* is *disproven* for both subreddits.

User-Based Analysis Two different results could be discovered by observing the user-based analysis for unique and permanent active users. First, the analysis of unique users classified as toxic shows a similar trend to the previous analyses. That demonstrates that r/relationships had fewer unique toxic users since 2018, while in r/relationship_advice the number of toxic users increased and stabilised "after" a while. This supports the argument that moderation was stricter for such users. Thus, new users are obligated to follow this social norm and not get banned. Second, a similar trend can also be seen among the permanent active users, where the toxic behaviour in r/relationships remained minimal or almost constant "after" the change. With respect to r/relationship_advice, it became more toxic "after" the change but stayed constant with only minimal changes. It is assumed that in both subreddits, the users did not change their toxic behaviour because of the design but simply continued because they already knew the moderation rules and never saw the design change as a huge impact. It is interesting to note, however, that this time, the toxic behaviour of permanently active users was usually higher with r/relationships "after" the change than with r/relationship_advice. Even so, there were fewer permanent active users for r/relationship_advice and significantly more for r/relationships.

Similar to the hypothesis *H1*, which was disproven in the discussion of the content-based analysis, the hypotheses *H2* and *H3* are also *disproven*.

Submission-Based Analysis In the submission-based analysis, the results showed that r/relationship_advice was more toxic to the same submissions than r/relationships until 2019. Whereas from 2020 onwards, users in r/relationships reacted more toxic to the same submissions. It is difficult to determine which subreddit was more toxic based on the method that was used. Additionally, it is not possible to state whether the design or the moderation was the reason for this trend. This variation in toxicity over the years ("before" and "after" the change) might be due to the fact that only the first posted submission was kept, and the remaining duplicates were not considered. It could be possible that comments on the other submissions were more commented on, perhaps more toxic, or vice versa.

Discussion of the Hypothesis H4 Based on the discussions on the content-, user- and submission-based analysis, the fourth hypothesis *(H4)* can now be answered. Apart from the results of the submission and permanent active user-based analysis, everything was aligned with this hypothesis. Thus, the results of Messerschmidt et al. (2023), on the subreddits with different degrees of disorder with regards to *H4*, were *confirmed*.

Now that all hypotheses have been discussed, the research question will be answered. The toxicity trend in r/relationship_advice was low "before" the change and only became more drastic "after" the change, whereas with r/relationships it went in the complete opposite direction. Thus, it can be interpreted that because Reddit itself had a design change, not everyone had less toxic behaviour, which was the case with r/relationship_advice. In this case, moderation played a huge role in controlling or reducing the toxic behaviour. The redesign from disordered to less disordered may have a minimal effect, clearly seen in the number of new users.

Finally, it could be said that the answer for the *RQ: "Does perceived disorder in an online environment increase toxicity?*" is clearly *no*. It can not be applied to the case of these two subreddits.

5.3 Theoretical and Practical Insights

Even if this thesis could not confirm the BWT in the online environments by verifying the hypotheses and *RQ*, quite a few theoretical and practical insights could be gained. First of all, this work extends the currently existing literature in the area of antisocial commenting behaviour in an online environment (Cheng et al., 2017; Del Vigna12 et al., 2017; S. Kumar et al., 2018). More specifically, this study provides some evidence on how to avoid antisocial behaviour, which was not fully related to the design of the online environment. Second, this work extends the research in the area of BWT in an online environment, where less literature is currently available. To the authors' knowledge, this work is the second work (following Messerschmidt et al. (2023)) where the focus is on commenting behaviour regarding toxicity related to the design of an online environment. Additionally, regarding the BWT in the online environment, it is the first study validating that this theory does not work in the online environment by using the platform Reddit. All in all, this study can also help future researchers to investigate the BWT in the online environment by looking at an online environment, where the moderation rules should be the same or similar. Regarding Reddit, it should be subreddits with the same or similar moderation rules.

In addition to the theoretical insight, practical insight could also be found as the results of this thesis. This thesis suggests that reducing antisocial behaviour on Reddit or other social media platforms can be investigated by focusing more on the moderation of the platforms. It is important to develop moderation rules that are clear to the users and can be fully implemented by the moderator or website owners. Especially for the Reddit owners, this is important, as all moderation rules for each subreddit are self-constructed. With this, it is possible to reduce toxic behaviour and control each environment (e.g. subreddit) separately. To achieve this, it is important to train the moderators intensively or to develop AI-based moderators that can detect if the users are breaking the rules or not. Also, it is essential to prevent this antisocial toxic behaviour as it has many negative side effects. These can be, for example, mental and emotional stress for the user (Duggan, 2017; Soares et al., 2023), reduction of online participation since users might feel uncomfortable being online (Sobieraj, 2018), affecting users' emotional feelings (Duggan, 2017), their reputation or personal safety (Duggan, 2017). Especially for underage users where the effect of antisocial behaviour like cyber-bullying could also be dangerous in a psychological view, such as the development of negative affective disorders, loneliness, anxiety, depression, and suicidal ideation (Nixon, 2014). Altogether, these ideas are critical to creating an online environment, which should be a much safer place by employing better and more concise moderation rules.

5.4 Limitations and Future Work

A few obstacles and limitations occurred during the research phase of this thesis. One of the main downsides of this research was the dataset itself. Since only one month is analysed for each year (instead of the full year) between 2016-2022, the time series plots could not give a full insight into the commenting behaviour between the defined time range. Furthermore, the toxicity analysis would take much more time since only 30k comments per run could be uploaded in the Communalytics free research version. Each run might take about 7-8 hours. Another problem was also that the data set in r/relationship_advice had data leaks in mid-November. A reason for this could be that on this time period the scrapping tool could not scrape the data, which kind of manipulated the data itself for that year. A general conclusion concerning the research question could not be realised, even if the subreddit research was detailed and those subreddit pairs had mostly the same metrics. More specifically, the diversity in moderation rules had a huge impact on the dataset obtained from both subreddits. Even if the results are true of the selected subreddits, it does not mean this commenting behaviour is true for all of Reddit. Especially for this platform with diverse subreddits as well as different moderation rules and topics.

This shows that there are many possibilities to expand the research of BWT in online environments. For example, to better investigate the BWT in an online environment, a platform should be explored, in which the moderation rules should be more consistent for the whole platform. Regarding the data set, it is better to focus on a shorter time range, for instance, 2-3 years. However, instead of looking at one month in each year, the focus could be on the whole year. Another essential point to be mentioned is the ability to use the premium version of Communalytic. This is indeed more expensive but allows the researcher to provide more data as input for the analyses. As an

alternative, it is also recommended to apply the toxicity analysis with Perspective API on a server with GPU power by using R and Python to get faster toxicity score results. Another idea would be to use other toxicity analysis approaches like Detoxify (Hanu & Unitary, 2020), which could give more accurate predictions of whether a comment is toxic or not. In addition to the methods, topic modelling could also be applied to better understand the content of the comments themself, that is, if the users are talking about the topic or being toxic.

6 Conclusion

The focus of this work was on finding out if the BWT is also applicable in online environments. More specifically, the aim was to discover if a redesign from a disordered to a less disordered website could decrease the toxic commenting behaviour. To achieve this, commenting behaviour with a "before" and "after" comparison has been conducted on the Reddit platform. To accomplish this, the data from the subreddits r/relationships and r/relationship_advice between 2016 and 2022 for each November were collected and analysed. The analysis was done in descriptive, content-based, user-based, and submission-based manners. For these methods, different approaches, such as sentiment, toxicity, and time series analysis with change point detection have been applied. In conclusion, the design component had no major impact on the commenting behaviour, while the moderation of each subreddit might play a higher role. The subreddits with stricter moderation rules also had less toxic comments, while the subreddit with less strict moderation had a higher toxicity. This inconvenient result, however, helps in understanding how toxic behaviour can be reduced or even prevented. Furthermore, this research expands the literature of the BWT in an online environment, which also opens the door for more opportunities for future research.

References

1. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker Jr, J. F. (2010). Detecting fake websites: The contribution of statistical learning theory. *Mis Quarterly*, 435–461.
2. Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. (2020). Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis. *Algorithms*, *13*(12), 345.
3. Aljafari, D. (2019). *Examining the effects of parasocial interaction and identification with social media influencers on collaborating brands* (Unpublished master's thesis).
4. Alsudani, F., & Casey, M. (2009). The effect of aesthetics on web credibility. *People and Computers XXIII Celebrating People and Technology*, 512–519.
5. Alsuliman, F., Bhattacharyya, S., Slhoub, K., Nur, N., & Chambers, C. N. (2022). Social media vs. news platforms: A cross-analysis for fake news detection using web scraping and nlp. In *Proceedings of the 15th international conference on pervasive technologies related to assistive environments* (pp. 190–196).
6. Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, *51*(2), 339–367.
7. Anvaka. (2020). *Anvaka/sayit: Visualization of related subreddits.* Retrieved from https://github.com/anvaka/sayit#the-data (Accessed 06.09.2023)
8. Aragón, P., Gómez, V., & Kaltenbrunner, A. (2017). Detecting platform effects in online discussions. *Policy & Internet*, *9*(4), 420–443.
9. Archive, I. (2023a). *Reddit front-page 2017.* Retrieved from https://web.archive.org/web/20170607000600/https://www.reddit.com/ (Accessed 05.08.2023)
10. Archive, I. (2023b). *Reddit front-page 2018.* Retrieved from https://web.archive.org/web/20180731231932/https://www.reddit.com/ (Accessed 05.08.2023)
11. Archive, I. (2023c). *r/relationshipadvice frontpage 2017.* Retrieved from https://web.archive.org/web/20171121183250/https://www.reddit.com/r/relationship_advice/ (Accessed 25.08.2023)
12. Archive, I. (2023d). *r/relationshipadvice frontpage 2018.* Retrieved from https://web.archive.org/web/20181115065045/https://www.reddit.com/r/relationship_advice/ (Accessed 25.08.2023)
13. Archive, I. (2023e). *r/relationshipadvice frontpage 2022.* Retrieved from https://web.archive.org/web/20221109041655/https://www.reddit.com/r/relationship_advice/ (Accessed 25.08.2023)
14. Archive, I. (2023f). *r/relationships frontpage 2017.* Retrieved from https://web.archive.org/web/20171115214042/https://www.reddit.com/r/relationships/ (Accessed 25.08.2023)
15. Archive, I. (2023g). *r/relationships frontpage 2018.* Retrieved from https://web.archive.org/web/20181120064342/https://www.reddit.com/r/relationships (Accessed 25.08.2023)
16. Archive, I. (2023h). *r/relationships frontpage 2022.* Retrieved from https://web.archive

.org/web/20221114035040/https://www.reddit.com/r/relationships/ (Accessed 25.08.2023)

17. Archive, I. (2024). *r/relationships faq page*. Retrieved from https://web.archive.org/web/20181120012309/https://www.reddit.com/r/relationships/wiki/index/ (Accessed 24.02.20224)
18. Archive.org. (2024). *Archive.org help page*. Retrieved from https://help.archive.org/help/archive-org-page-overview/ (Accessed 04.02.2024)
19. Atari, M., Davani, A. M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., & Dehghani, M. (2022). Morally homogeneous networks and radicalism. *Social Psychological and Personality Science*, *13*(6), 999–1009.
20. B0tRank. (2023). *B0trank*. Retrieved from https://botrank.pastimes.eu/ (Accessed 15.09.2023)
21. Barlett, C. P., Gentile, D. A., & Chew, C. (2016). Predicting cyberbullying from anonymity. *Psychology of Popular Media Culture*, *5*(2), 171.
22. Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application* (Vol. 104). prentice Hall Englewood Cliffs.
23. Bauerly, M., & Liu, Y. (2008). Effects of symmetry and number of compositional elements on interface and design aesthetics. *Intl. Journal of Human–Computer Interaction*, *24*(3), 275–287.
24. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 830–839).
25. Blumstein, A. (1995). Youth violence, guns, and the illicit-drug industry. *J. Crim. L. & Criminology*, *86*, 10.
26. Bonnardel, N., Piolat, A., & Le Bigot, L. (2011). The impact of colour on website appeal and users' cognitive processes. *Displays*, *32*(2), 69–80.
27. Borah, P. (2014). Does it matter where you read the news story? interaction of incivility and news frames in the political blogosphere. *Communication Research*, *41*(6), 809–827.
28. Bottomley, P. A., & Doyle, J. R. (2006). The interactive effects of colors and products on perceptions of brand logo appropriateness. *Marketing Theory*, *6*(1), 63–83.
29. Bowling, B. (1999). The rise and fall of new york murder: zero tolerance or crack's decline? *British Journal of Criminology*, *39*(4), 531–554.
30. Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference ranlp-2009* (pp. 50–54).
31. Brown, N. J., & Coyne, J. C. (2018). Does twitter language reliably predict heart disease? a commentary on eichstaedt et al.(2015a). *PeerJ*, *6*, e5656.
32. Browning, C. R., Soller, B., Gardner, M., & Brooks-Gunn, J. (2013). "feeling disorder" as a comparative and contingent process: Gender, neighborhood conditions, and adolescent mental health. *Journal of health and social behavior*, *54*(3), 296–314.
33. Burnett, G. (2009). Colliding norms, community, and the place of online information: The case of archive. org. *Library Trends*, *57*(4), 694–710.

34. Bursik, R. J., & Grasmick, H. G. (1993). The dimensions of effective community control. *Lexington, MA: Lexington Books*.
35. Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., & Ceulemans, E. (2017). Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. *Behavior research methods*, *49*, 988–1005.
36. Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic journal of communication*, *23*(1), 46–65.
37. Cerdá, M., Messner, S. F., Tracy, M., Vlahov, D., Goldmann, E., Tardiff, K. J., & Galea, S. (2010). Investigating the effect of social changes on age-specific gun-related homicide rates in new york city during the 1990s. *American journal of public health*, *100*(6), 1107–1115.
38. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 acm on web science conference* (pp. 13–22).
39. Chauhan, P., Cerdá, M., Messner, S. F., Tracy, M., Tardiff, K., & Galea, S. (2011). Race/ethnic-specific homicide rates in new york city: evaluating the impact of broken windows policing and crack cocaine markets. *Homicide studies*, *15*(3), 268–290.
40. Chen, J., & Gupta, A. K. (2012). Parametric statistical change point analysis: with applications to genetics, medicine, and finance.
41. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *2012 international conference on privacy, security, risk and trust and 2012 international confernece on social computing* (pp. 71–80).
42. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (pp. 1217–1230).
43. Chick, K. (2020). Harmful comments on social media. *York L. Rev.*, *1*, 83.
44. Churruca, K., Ellis, L. A., & Braithwaite, J. (2018). 'broken hospital windows': debating the theory of spreading disorder and its application to healthcare organizations. *BMC health services research*, *18*, 1–6.
45. Cohen, D., Spear, S., Scribner, R., Kissinger, P., Mason, K., & Wildgen, J. (2000). " broken windows" and the risk of gonorrhea. *American journal of public health*, *90*(2), 230.
46. Cohen, K. B. (2014). Chapter 6-biomedical natural language processing and text mining. *Methods in biomedical informatics*, 141–177.
47. Collings-Wells, S. (2022). From black power to broken windows: Liberal philanthropy and the carceral state. *Journal of Urban History*, *48*(4), 739–759.
48. Corman, H., & Mocan, N. (2005). Carrots, sticks, and broken windows. *The Journal of Law and Economics*, *48*(1), 235–266.
49. Cyr, D. (2008). Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of management information systems*, *24*(4), 47–72.
50. Cyr, D., Head, M., & Larios, H. (2010). Colour appeal in website design within and across cultures: A multi-method evaluation. *International journal of human-computer studies*,

68(1-2), 1–21.

51. Dang, V. T. (2021). Social networking site involvement and social life satisfaction: The moderating role of information sharing. *Internet Research*, *31*(1), 80–99.
52. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 512–515).
53. Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first italian conference on cybersecurity (itasec17)* (pp. 86–95).
54. DiIulio Jr, J. J. (1995). Arresting ideas: Tougher law enforcement is driving down urban crime. *Policy Review*(74), 12–17.
55. Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the international aaai conference on web and social media* (Vol. 5, pp. 11–17).
56. Ding, Y., & Qiu, L. (2017). The impact of celebrity-following activities on endorsement effectiveness on microblogging platforms: A parasocial interaction perspective. *Nankai Business Review International*, *8*(2), 158–173.
57. Duggan, M. (2017). Online harassment 2017.
58. Eck, J. E., & Maguire, E. R. (2000). Have changes in policing reduced violent crime? an assessment of the evidence. *The crime drop in America*, *207*, 207–265.
59. Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 122, p. 16).
60. Ellis, L. A., Churruca, K., Tran, Y., Long, J. C., Pomare, C., & Braithwaite, J. (2020). An empirical application of "broken windows" and related theories in healthcare: examining disorder, patient safety, staff outcomes, and collective efficacy in hospitals. *BMC health services research*, *20*(1), 1–12.
61. Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, *45*(1), 1–34.
62. Fabian, B., Baumann, A., & Keil, M. (2015). Privacy on reddit? towards large-scale user classification. In *Ecis*.
63. Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., ... Al-qaness, M. A. (2021). Social media toxicity classification using deep learning: real-world application uk brexit. *Electronics*, *10*(11), 1332.
64. Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th acm conference on web science* (pp. 87–96).
65. Fernando, A., Feldmann, L., & Stockhausen, P. (2023). *A review on digital parasocial relationships*.
66. Flury, C. G. (2017). Social media as a leadership tool for nurse executives: Claiming the corner office. *Nursing Economics*, *35*(5), 272–276.

67. Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *Chi'03 extended abstracts on human factors in computing systems* (pp. 722–723).
68. Fredheim, R., Moore, A., & Naughton, J. (2015). Anonymity and online commenting: The broken windows effect and the end of drive-by commenting. In *Proceedings of the acm web science conference* (pp. 1–8).
69. Gau, J. M., Corsaro, N., & Brunson, R. K. (2014). Revisiting broken windows theory: A test of the mediation impact of social mechanisms on the disorder–fear relationship. *Journal of Criminal Justice*, *42*(6), 579–588.
70. Gau, J. M., & Pratt, T. C. (2008). Broken windows or window dressing? citizens'(in) ability to tell the difference between disorder and crime. *Criminology & Public Policy*, *7*(2), 163–194.
71. Gau, J. M., & Pratt, T. C. (2010). Revisiting broken windows theory: Examining the sources of the discriminant validity of perceived disorder and crime. *Journal of criminal justice*, *38*(4), 758–766.
72. Gil, Y., Chai, Y., Gorodissky, O., & Berant, J. (2019). White-to-black: Efficient distillation of black-box adversarial attacks. *arXiv preprint arXiv:1904.02405*.
73. Google Jigsaw. (2023a). *About the perspective api.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/about-the-api-faqs?language=en_US (Accessed 15.12.2023)
74. Google Jigsaw. (2023b). *About the perspective api: Attributes and languages.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US (Accessed 15.12.2023)
75. Google Jigsaw. (2023c). *About the perspective api: Model cards.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US (Accessed 15.09.2023)
76. Google Jigsaw. (2023d). *About the perspective api: Score.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US (Accessed 15.09.2023)
77. Google Jigsaw. (2023e). *About the perspective api: Training data.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US (Accessed 15.09.2023)
78. Google Jigsaw. (2023f). *Perspective api: Contact support.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/contactsupport?language=en_US (Accessed 15.12.2023)
79. Google Jigsaw. (2023g). *Perspective api documentation: Getting started.* Perspective API. Retrieved from https://developers.perspectiveapi.com/s/docs-get-started?language=en_US (Accessed 15.12.2023)
80. Google Jigsaw. (2023h). *Using machine learning to reduce toxicity online.* Perspective API. Retrieved from https://perspectiveapi.com/ (Accessed 15.12.2023)
81. Grimes, M., Marquardson, J., & Nunamaker, J. (2014). Broken windows, bad passwords: Influencing secure user behavior via website design.

82. Gruzd, A., & Mai, P. (2022). *Communalytic: A research tool for studying online communities and online discourse.*
83. Gruzd, A., Mai, P., & Soares, F. B. (2023). From trolling to cyberbullying: Using machine learning and network analysis to study anti-social behavior on social media. In *Proceedings of the 34th acm conference on hypertext and social media* (pp. 1–2).
84. Gruzd, A., Mai, P., & Vahedi, Z. (2020). Studying anti-social behaviour on reddit with communalytic.
85. Hammami, B., & Afram, M. (2022). *Analysis and evaluation ofvisuospatial complexity models.*
86. Hanu, L., & Unitary, t. (2020, November). *Detoxify*. Retrieved from https://github.com/unitaryai/detoxify doi: 10.5281/zenodo.7925667
87. Harcourt, B. E. (2005). *Illusion of order: The false promise of broken windows policing*. Harvard University Press.
88. Harcourt, B. E., & Ludwig, J. (2006). Broken windows: New evidence from new york city and a five-city social experiment. *U. Chi. L. Rev.*, *73*, 271.
89. Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing arabic text on social media. *Heliyon*, *7*(2).
90. Henwood, A., Rinck, M., & Krpan, D. (2023). Pandemic related changes in social interaction are associated with changes in automatic approach-avoidance behaviour. *Scientific Reports*, *13*(1), 4637.
91. Hessel, J., Tan, C., & Lee, L. (2016). Science, askscience, and badscience: On the coexistence of highly related communities. In *Proceedings of the international aaai conference on web and social media* (Vol. 10, pp. 171–180).
92. Hiaeshutter-Rice, D., & Hawkins, I. (2022). The language of extremism on social media: An examination of posts, comments, and themes on reddit. *Frontiers in Political Science*, *4*, 805008.
93. Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–24.
94. Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human communication research*, *41*(4), 557–576.
95. Hua, Y., Ristenpart, T., & Naaman, M. (2020). Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 272–282).
96. Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).
97. Hynes, N. (2009). Colour and meaning in corporate logos: An empirical study. *Journal of Brand Management*, *16*, 545–555.
98. Icon-Icons. (2024). *Reddit icon source.* Retrieved from https://icon-icons.com/de/

symbol/Reddit/121349 (Accessed 04.02.2024)

99. Internet Archive. (1996). *Internet archive: Digital library of free & borrowable books, movies, music & wayback machine.* Retrieved from https://archive.org/ (Accessed 06.08.2023)
100. Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., & Lecheler, S. (2022). Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality. *Journal of Computer-Mediated Communication*, *27*(1), zmab019.
101. Jain, E., Brown, S., Chen, J., Neaton, E., Baidas, M., Dong, Z., ... Artan, N. S. (2018). Adversarial text generation for google's perspective api. In *2018 international conference on computational science and computational intelligence (csci)* (pp. 1136–1141).
102. James, N., Zhang, W., & Matteson, D. (2023, Jul). *Package 'ecp' - the comprehensive r archive network.* Retrieved from https://cran.r-project.org/web/packages/ecp/ecp.pdf (Accessed 20.12.2023)
103. James, N. A., & Matteson, D. S. (2013). ecp: An r package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*.
104. Joanes, A. (1999). Does the new york city police department deserve credit for the decline in new yorks city's homicide rates-a cross-city comparison of policing strategies and homicide rates. *Colum. JL & Soc. Probs.*, *33*, 265.
105. Jusoh, S., & Al-Fawareh, H. M. (2007). Natural language interface for online sales systems. In *2007 international conference on intelligent and advanced systems* (pp. 224–228).
106. Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, *53*(1), 59–68.
107. Kargaran, A. H. (2021). *Visual clutter python library.* https://github.com/kargaranamir/visual-clutter. GitHub. (Accessed 10.09.2023)
108. Kawahara, Y., & Sugiyama, M. (2012, apr). Sequential change-point detection based on direct density-ratio estimation. *Stat. Anal. Data Min.*, *5*(2), 114–127. Retrieved from https://doi.org/10.1002/sam.10124 doi: 10.1002/sam.10124
109. Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *science*, *322*(5908), 1681–1685.
110. Kelling, G. L., & Bratton, W. J. (1997). Declining crime rates: Insiders' views of the new york city story. *J. crim. L. & criminology*, *88*, 1217.
111. Kelling, G. L., & Coles, C. M. (1997). *Fixing broken windows: Restoring order and reducing crime in our communities*. Simon and Schuster.
112. Kelling, G. L., & Sousa, W. H. (2001). *Do police matter?: An analysis of the impact of new york city's police reforms*. CCI Center for Civic Innovation at the Manhattan Institute.
113. Kelling, G. L., Wilson, J. Q., et al. (1982). Broken windows. *Atlantic monthly*, *249*(3), 29–38.
114. KeyserSosa. (2023, Apr). *An update regarding reddit's api.* reddit. Retrieved from https://www.reddit.com/r/reddit/comments/12qwagm/an_update_regarding_reddits_api/ ([Online forum post], Accessed 06.09.2023)
115. Khanaferov, D., Luc, C., & Wang, T. (2014). Social network data mining using natural language processing and density based clustering. In *2014 ieee international conference on semantic*

computing (pp. 250–251).

116. Kim, T., & Wurste, K. (2023). *Emoji*. Retrieved from https://pypi.org/project/emoji/ (Accessed 10.11.2023)
117. Kitchens, B., Johnson, S. L., & Gray, P. (2020). Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, *44*(4).
118. Koh, M. K. (2018). *The 15 best discussions on reddit ever*.
119. Koutamanis, M., Vossen, H. G., & Valkenburg, P. M. (2015). Adolescents' comments in social media: Why do adolescents receive negative feedback and who is most at risk? *Computers in Human Behavior*, *53*, 486–494.
120. Kumar, D., Hancock, J., Thomas, K., & Durumeric, Z. (2023). Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the acm web conference 2023* (pp. 2797–2807).
121. Kumar, S., Cheng, J., & Leskovec, J. (2017). Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th international conference on world wide web companion* (pp. 947–950).
122. Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference* (pp. 933–943).
123. Lee, M. J., & Chun, J. W. (2016). Reading others' comments and public opinion poll results on social media: Social judgment and spiral of empowerment. *Computers in Human Behavior*, *65*, 479–487.
124. Lee, Y., & Kozar, K. A. (2006). Investigating the effect of website quality on e-business success: An analytic hierarchy process (ahp) approach. *Decision support systems*, *42*(3), 1383–1401.
125. Leitner, L. (2020, Apr). *Redditcleanerlorenz leitner*. Retrieved from https://pypi.org/project/redditcleaner/ (Accessed 10.11.2023)
126. Liao, S. (2018, Apr). *Reddit begins rolling out first redesign in a decade*. The Verge. Retrieved from https://www.theverge.com/2018/4/2/17190244/reddit-redesign-begins-rolling-out (Accessed 06.08.2023)
127. Link, N. W., Kelly, J. M., Pitts, J. R., Waltman-Spreha, K., & Taylor, R. B. (2017). Reversing broken windows: Evidence of lagged, multilevel impacts of risk perceptions on perceptions of incivility. *Crime & Delinquency*, *63*(6), 659–682.
128. Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.
129. Liu, J., Wu, J. S., & Che, T. (2019). Understanding perceived environment quality in affecting tourists' environmentally responsible behaviours: A broken windows theory perspective. *Tourism Management Perspectives*, *31*, 236–244.
130. Loten, A. (2018, Apr). *Reddit ceo revamped outdated website from the it foundations*. THE WALL STREET JOURNAL. Retrieved from https://www.wsj.com/articles/reddit-ceo-revamped-outdated-website-from-the-it-foundations-1523396124 (Accessed 06.08.2023)
131. Lowry, P. B., Moody, G. D., & Chatterjee, S. (2017). Using the control balance theory to explain social media deviance. In *Hawaii international conference on system sciences (hicss-50),*

big island, hi, january (pp. 4–7).

132. Lowry, P. B., Wilson, D. W., & Haig, W. L. (2014). A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust. *International Journal of Human-Computer Interaction*, *30*(1), 63–93.
133. Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, *156*(4), 133–162.
134. Marotti, A. (2018). Reddit to open chicago office as part of advertising push. *Chicagotribune. Com*.
135. Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015). Toxicity detection in multiplayer online games. In *2015 international workshop on network and systems support for games (netgames)* (pp. 1–6).
136. Marwick, A. E. (2017). Scandal or sex crime? gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, *19*, 177–191.
137. Maskaly, J., & Boggess, L. N. (2014). Broken windows theory. *The Encyclopedia of Theoretical Criminology*, 1–4.
138. Massanari, A. (2017). # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New media & society*, *19*(3), 329–346.
139. Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media+ Society*, *5*(2), 2056305119836778.
140. Matteson, D. S., & James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, *109*(505), 334–345.
141. McLaughlin, A. (2018, May). *Reddit's head of design on how the site was redesigned.* Design Week. Retrieved from https://www.designweek.co.uk/issues/9-15-april-2018/reddits-head-design-site-redesigned/ (Accessed 06.09.2023)
142. Medvedev, A. N., Lambiotte, R., & Delvenne, J.-C. (2019). The anatomy of reddit: An overview of academic research. *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*, 183–204.
143. Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505–1512.
144. Messerschmidt, N., Gundlach, J., & Baumann, A. (2023). A look through a broken window: The relationship between disorder and toxicity on social networking sites.
145. Mills, K. A., & Chandra, V. (2011). Microblogging as a literacy practice for educational communities. *Journal of Adolescent & Adult Literacy*, *55*(1), 35–45.
146. Mills, R. A. (2018). Pop-up political advocacy communities on reddit. com: Sandersforpresident and the donald. *Ai & Society*, *33*, 39–54.
147. Mintz, A. P. (2002). *Web of deception: Misinformation on the internet*. Information Today, Inc.
148. Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2020). "and we will fight for

our race!" a measurement study of genetic testing conversations on reddit and 4chan. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 452–463).

149. Neutatz, F., Chen, B., Abedjan, Z., & Wu, E. (2021). From cleaning before ml to cleaning for ml. *IEEE Data Eng. Bull.*, *44*(1), 24–41.
150. Nixon, C. L. (2014). Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent health, medicine and therapeutics*, 143–158.
151. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).
152. Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., & Giordano, S. (2023). Toxic bias: Perspective api misreads german as more toxic. *arXiv preprint arXiv:2312.12651*.
153. Nudd, T. (2014, Dec). *The meaning of 35 brand names, from etsy to reddit.* ADWEEK. Retrieved from https://www.adweek.com/creativity/meaning-35-brand-names-etsy-reddit-161694/ (Accessed 15.12.2023)
154. Obadimu, A., Mead, E., Hussain, M. N., & Agarwal, N. (2019). Identifying toxicity within youtube video comment. In *Social, cultural, and behavioral modeling: 12th international conference, sbp-brims 2019, washington, dc, usa, july 9–12, 2019, proceedings 12* (pp. 214–223).
155. O'Brien, D. T., Farrell, C., & Welsh, B. C. (2019a). Broken (windows) theory: A meta-analysis of the evidence for the pathways from neighborhood disorder to resident health outcomes and behaviors. *Social science & medicine*, *228*, 272–292.
156. O'Brien, D. T., Farrell, C., & Welsh, B. C. (2019b). Looking through broken windows: The impact of neighborhood disorder on aggression and fear of crime is an artifact of research design. *Annual Review of Criminology*, *2*, 53–71.
157. OldReddit. (2023). *Reddit submission page 2017.* Retrieved from https://old.reddit.com/r/Divorce/comments/15j4odd/angry_fck_the_partiarchy_wife/ (Accessed 05.08.2023)
158. Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in facebook and its application to e-learning. *Computers in human behavior*, *31*, 527–541.
159. O'Brien, D. T., & Sampson, R. J. (2015). Public and private spheres of neighborhood disorder: Assessing pathways to violence using large-scale digital records. *Journal of research in crime and delinquency*, *52*(4), 486–510.
160. Pardes, A. (2018, Apr). *The inside story of reddit's redesign.* Wired. Retrieved from https://www.wired.com/story/reddit-redesign/ (Accessed 06.08.2023)
161. Peebles, E. (2014). Cyberbullying: Hiding behind the screen. *Paediatrics & child health*, *19*(10), 527–528.
162. Pellert, M., Metzler, H., Matzenberger, M., & Garcia, D. (2022). Validating daily social media macroscopes of emotions. *Scientific reports*, *12*(1), 11236.
163. Plank, S. B., Bradshaw, C. P., & Young, H. (2009). An application of "broken-windows" and related theories to the study of disorder, fear, and collective efficacy in schools. *American*

Journal of Education, *115*(2), 227–247.

164. Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, *7*(2), 20563051211019004.
165. Qu, H., Nascimento, M. S., Qomariyah, N. N., & Kazakov, D. L. (2016). Integrating time series with social media data in an ontology for the modelling of extreme financial events. In *International conference on language resources and evaluation* (pp. 57–63).
166. Rachele, J. N., Wood, L., Nathan, A., Giskes, K., & Turrell, G. (2016). Neighbourhood disadvantage and smoking: Examining the role of neighbourhood-level psychosocial characteristics. *Health & place*, *40*, 98–105.
167. Ramos, J., & Torgler, B. (2012). Are academics messy? testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics*, *8*(3), 563–577.
168. Reddit. (2020, Dec). *Reddit's 2020 year in review.* Author. Retrieved from https://www.redditinc.com/blog/reddits-2020-year-in-review/ (Accessed 15.12.2023)
169. Reddit. (2023a). *Reddit front-page 2023.* reddit. (Accessed 05.08.2023)
170. Reddit. (2023b). *Reddit submission page 2023.* Retrieved from https://www.reddit.com/r/Divorce/comments/15j4odd/angry_fck_the_partiarchy_wife/ (Accessed 05.08.2023)
171. Reddit. (2023c, Oct). *What are communities or "subreddits"? – reddit help.* reddit. Retrieved from https://support.reddithelp.com/hc/en-us/articles/204533569-What-are-communities-or-subreddits- (Accessed 15.12.2023)
172. Reddit. (2023d, Oct). *What are public, restricted, private, and premium-only ... - reddit help.* reddit. Retrieved from https://support.reddithelp.com/hc/en-us/articles/360060416112-What-are-public-restricted-private-and-premium-only-communities (Accessed 18.12.2023)
173. Roberts, D. E. (1998). Race, vagueness, and the social meaning of order-maintenance policing. *J. Crim. L. & Criminology*, *89*, 775.
174. Roose, K. (2021). The gamestop reckoning was a long time coming. *New York Times*, *28*.
175. Rosenfeld, R., Fornango, R., & Rengifo, A. F. (2007). The impact of order-maintenance policing on new york city homicide and robbery rates: 1988-2001. *Criminology*, *45*(2), 355–384.
176. Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of vision*, *7*(2), 17–17.
177. Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470.
178. Ross, C. E., & Mirowsky, J. (2009). Neighborhood disorder, subjective alienation, and distress. *Journal of health and social behavior*, *50*(1), 49–64.
179. Sachdev, D. (2023). Mil-std-1553 intrusion detection using ecp e-divisive algorithm.
180. Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American journal of sociology*, *105*(3), 603–651.

181. Sampson, R. J., & Raudenbush, S. W. (2001). *Disorder in urban neighborhoods: Does it lead to crime*. US Department of Justice, Office of Justice Programs, National Institute of ....
182. Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of "broken windows". *Social psychology quarterly*, *67*(4), 319–342.
183. Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *science*, *277*(5328), 918–924.
184. Savigny, J., & Purwarianti, A. (2017). Emotion classification on youtube comments using word embedding. In *2017 international conference on advanced informatics, concepts, theory, and applications (icaicta)* (pp. 1–5).
185. Savolainen, J. (2007). Public disorder and business victimization: Findings from a survey of female entrepreneurs. *Crime Prevention and Community Safety*, *9*, 1–20.
186. Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, *21*(7), 1417–1443.
187. Seife, C. (2014). *Virtual unreality: Just because the internet told you, how do you know it's true?* Penguin.
188. Shaw, C. R., & McKay, H. D. (1942). Juvenile delinquency and urban areas.
189. Shen, Q., & Rosé, C. P. (2022). A tale of two subreddits: Measuring the impacts of quarantines on political engagement on reddit. In *Proceedings of the international aaai conference on web and social media* (Vol. 16, pp. 932–943).
190. Sheth, A., Shalin, V. L., & Kursuncu, U. (2022). Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, *490*, 312–318.
191. Silver, E. R. (2021). Internet governance and online identity: Searching for correlations between user anonymity and user behaviour.
192. Silverman, E. B. (1999). *Nypd battles crime: Innovative strategies in policing*. UPNE.
193. Similarweb. (2023, Dec). *Top websites ranking - most visited websites in december 2023 - similarweb*. Author. Retrieved from https://www.similarweb.com/top-websites/ (Accessed 2023-12-15)
194. Singh, S. (2019). Everything in moderation: An analysis of how internet platforms are using artificial intelligence to moderate user-generated content. *New America*, *22*, 1–42.
195. Skogan, W. (2015). Disorder and decline: The state of research. *Journal of Research in Crime and Delinquency*, *52*(4), 464–485.
196. Skogan, W. G. (1990). Disorder and decline: Crime and the spiral of decay in american cities.
197. Skogan, W. G. (1992). *Disorder and decline: Crime and the spiral of decay in american neighborhoods*. Univ of California Press.
198. Slonje, R., Smith, P. K., & Frisén, A. (2012). Processes of cyberbullying, and feelings of remorse by bullies: A pilot study. *European Journal of Developmental Psychology*, *9*(2), 244–259.
199. Soares, F. B., Gruzd, A., Jacobson, J., & Hodson, J. (2023). To troll or not to troll: Young adults' anti-social behaviour on social media. *PLoS one*, *18*(5), e0284374.
200. Sobieraj, S. (2018). Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, *21*(11), 1700–1714.
201. Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk

radio, and cable news. *Political Communication*, *28*(1), 19–41.

202. Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 proceedings*.
203. Statista. (2024, Jan). *Social media - anzahl der nutzer weltweit bis 2023.* Retrieved from https://de.statista.com/statistik/daten/studie/739881/umfrage/monatlich-aktive-social-media-nutzer-weltweit/ (Online; accessed 2024-02-13)
204. stuck_in_the_matrix, Watchful1, & RaiderBDev. (2020). *Reddit comments/submissions 2005-06 to 2023-12.* Retrieved from https://academictorrents.com/details/9c263fc85366c1ef8f5bb9da0203f4c8c8db75f4 (Accessed 18.09.2023)
205. Subreddit-stats. (2023). *Subreddit stats.* Subreddit stats. Retrieved from https://subredditstats.com/ (Accessed 06.09.2023)
206. Tabassum, A., & Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, *7*(06), 4864–4867.
207. Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, *2*, 325–347.
208. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267–307.
209. Tait, A. (2020, Jul). *Who writes the crazy stories on reddit's r/relationships?* Vice. Retrieved from https://www.vice.com/en/article/4ay4vn/reddit-relationships-fake-stories-authors (Accessed 15.02.2024)
210. Taylor, R. B. (2001). Breaking away from broken windows: Evidence from baltimore neighborhoods and the nationwide fight against crime, grime, fear and decline. *New York, NY: Westvie*.
211. Thapa, B. (2022). Sentiment analysis of cybersecurity content on twitter and reddit. *arXiv preprint arXiv:2204.12267*.
212. Thelwall, M. (2014). Sentiment analysis and time series with twitter. *Twitter and society*, *1*.
213. Theocharopoulos, P. C., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., & Plagianakos, V. P. (2023). Analysing sentiment change detection of covid-19 tweets. *Neural Computing and Applications*, 1–11.
214. Trager, J., Ziabari, A. S., Davani, A. M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., ... others (2022). The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.
215. Vijayarani, S., Ilamathi, M. J., Nithya, M., et al. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.
216. Vilalta, C. J., Lopez, P., Fondevila, G., & Siordia, O. (2020). Testing broken windows theory in mexico city. *Social science quarterly*, *101*(2), 558–572.
217. Watchful1. (2023). *Pushshift dump utils.* GitHub. Retrieved from https://github.com/Watchful1/PushshiftDumps (Accessed 06.09.2023)

218. Welsh, B. C., Braga, A. A., & Bruinsma, G. J. (2015). Reimagining broken windows: From theory to policy. *Journal of Research in Crime and Delinquency*, *52*(4), 447–463.
219. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399).
220. Xia, Y., Zhu, H., Lu, T., Zhang, P., & Gu, N. (2020). Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, *4*(CSCW2), 1–23.
221. Xu, Y., Fiedler, M. L., & Flaming, K. H. (2005). Discovering the impact of community policing: The broken windows thesis, collective efficacy, and citizens' judgment. *Journal of Research in crime and Delinquency*, *42*(2), 147–186.
222. Yang, S. (2014). Social disorder and physical disorder at places. *Encyclopedia of Criminology and Criminal Justice*, *4922*, 4932.
223. Young, A. (2013). *Street art, public city: Law, crime and the urban imagination*. Routledge.
224. Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems*, *60*, 617–663.
225. Zadig, S. M., & Tejay, G. (2010). Securing is assets through hacker deterrence: A case study. In *2010 ecrime researchers summit* (pp. 1–7).
226. Zaheri, S., Leath, J., & Stroud, D. (2020). Toxic comment classification. *SMU Data Science Review*, *3*(1), 13.
227. Zhelev, Z., & Iliev, D. (2023). Opportunities and limitations of remote testing through online proctors. In *Developing curriculum for emergency remote learning environments* (pp. 100–120). IGI Global.
228. Zimbardo, P. G. (1973). A field experiment in auto shaping. *Vandalism*, 85–90.

7 Appendix

A

A.1 Reddit Front Page Design

Figures 18-20 display the front page design from Reddit for the years 2017, 2018, and 2023.

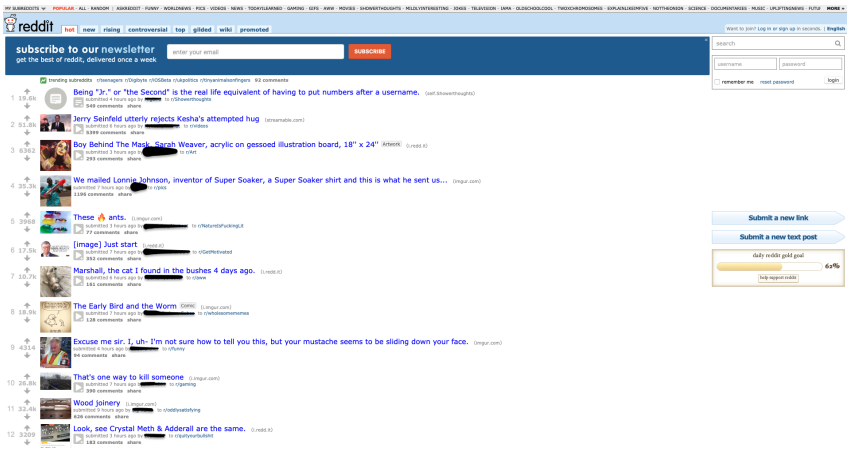

Figure 18: **Reddit front page design in 2017 (Archive, 2023a).**

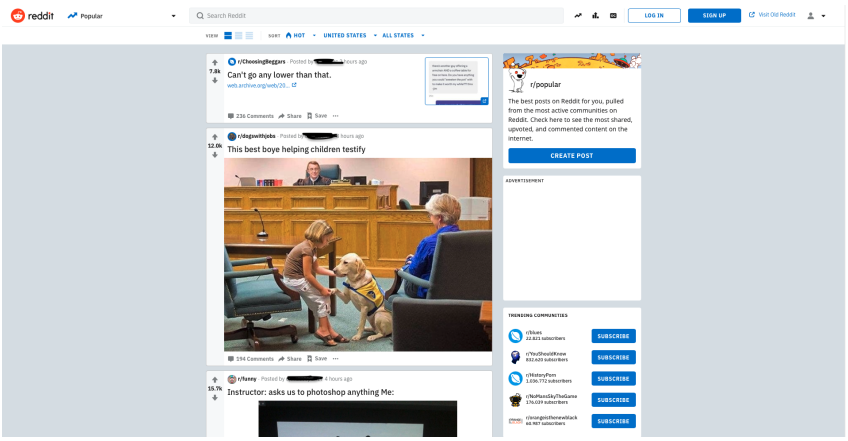

Figure 19: **Reddit front page design in 2018 (Archive, 2023b).**

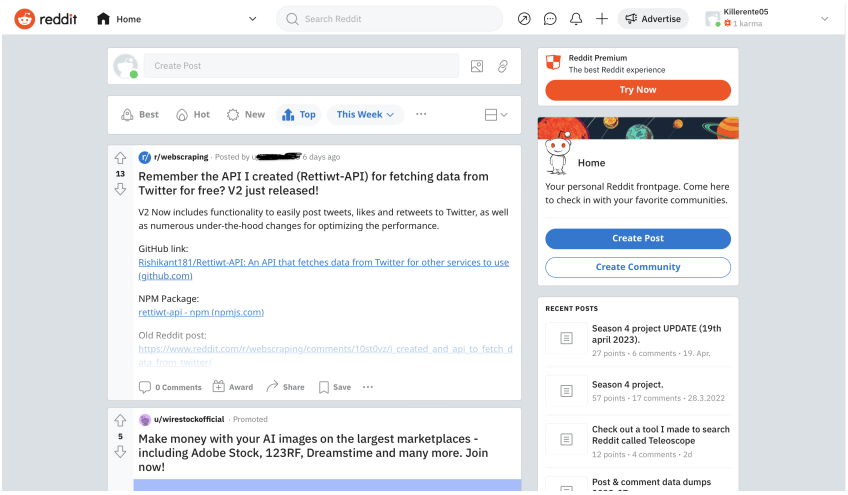

Figure 20: **Reddit front page design in 2023 (Reddit, 2023a).**

A.2 Reddit Submission and Commenting Section Design

Figures 21-24 show the design of the submission and commenting section from Reddit for the years 2017 and 2023.

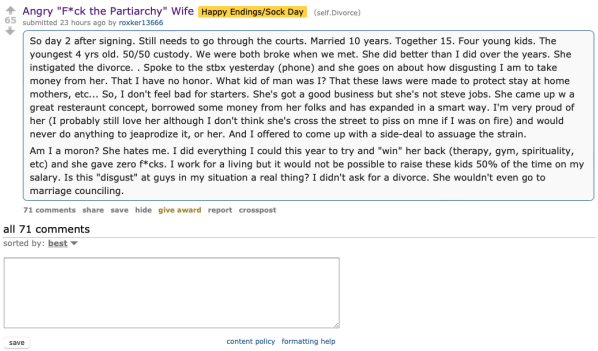

Figure 21: **Reddit submission design in 2017 (OldReddit, 2023).**

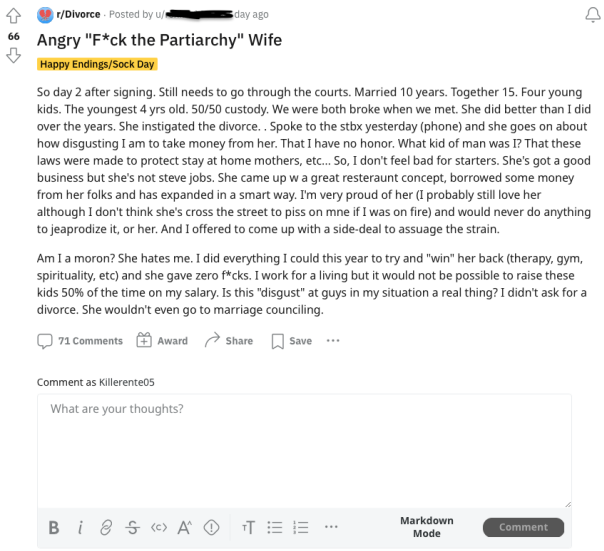

Figure 22: **Reddit submission design in 2023 (Reddit, 2023b).**

Figure 23: **Reddit comment section design in 2017 (OldReddit, 2023).**

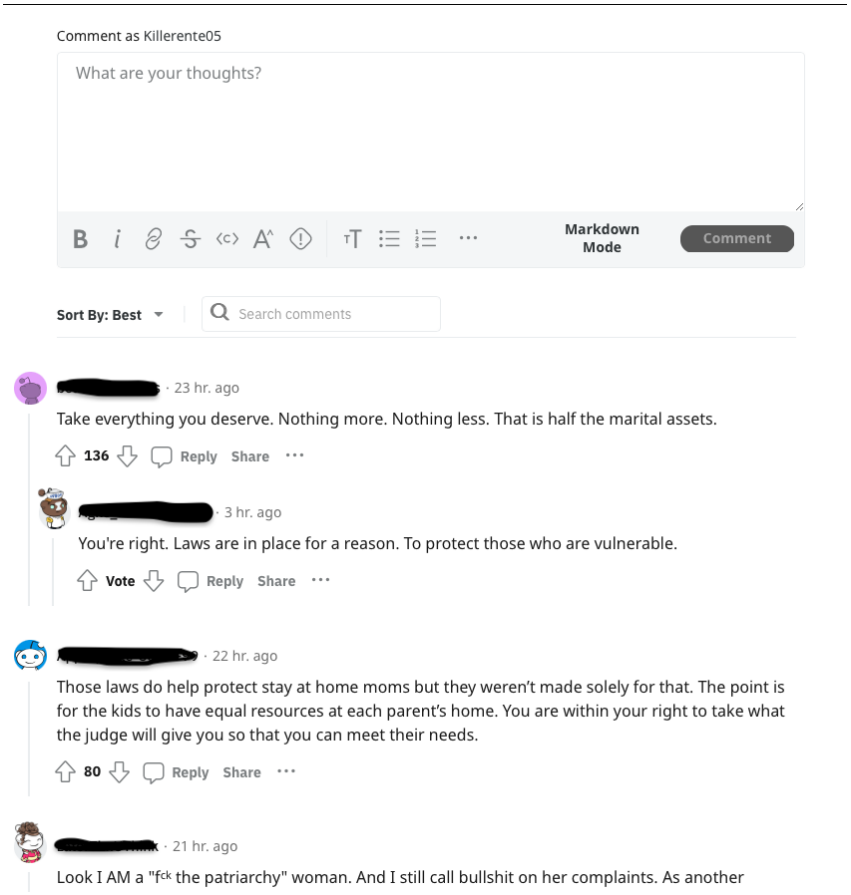

Figure 24: **Reddit comment section design in 2023 (Reddit, 2023b).**

B

B.1 Subreddit Pair Clutter Score

Table 11 illustrates the clutter score for each year's front page of the subreddit pairs.

Year	r/relationships	r/relationship_advice
2016	Fc:4.487/Se:3.338	Fc:4.683/Se:2.985
2017	Fc:4.607/Se:3.331	Fc:5.111/Se:3.169
2018	Fc:4.107/Se:2.765	Fc:4.355/Se:2.887
2019	Fc:4.000/Se:2.798	Fc:4.089/Se:2.638
2020	Fc:3.852/Se:2.761	Fc:3.953/Se:2.426
2021	Fc:4.058/Se:2.824	Fc:4.424/Se:2.637
2022	Fc:4.251/Se:3.045	Fc:4.973/Se:3.021

Table 11: **Overview of clutter score for front page of subreddit pair.**

C

C.1 Subreddit Pair Design

Figures 25 to 30 display the front page design for the subreddits r/relationships and r/relationship_advice for the years 2017, 2018 and 2022.

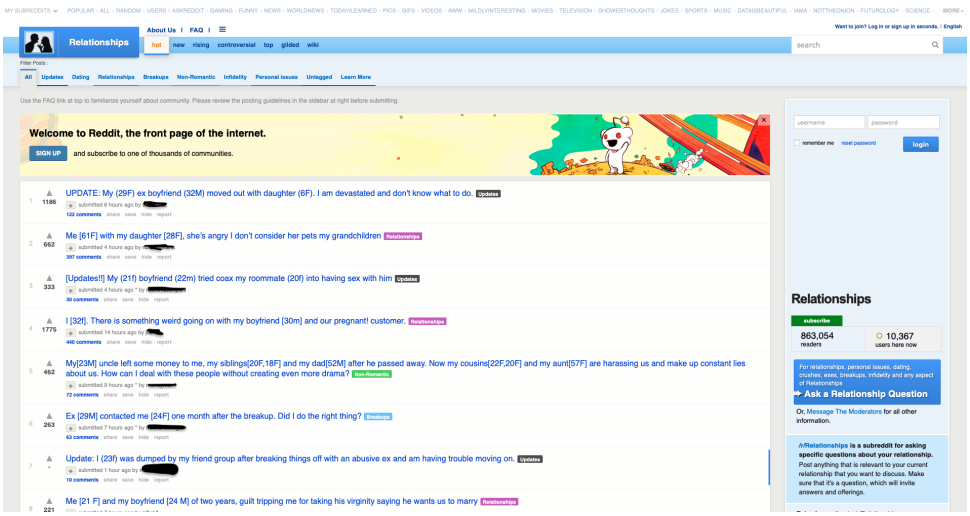

Figure 25: **Front page design of r/relationships in 2017 (Archive, 2023f).**

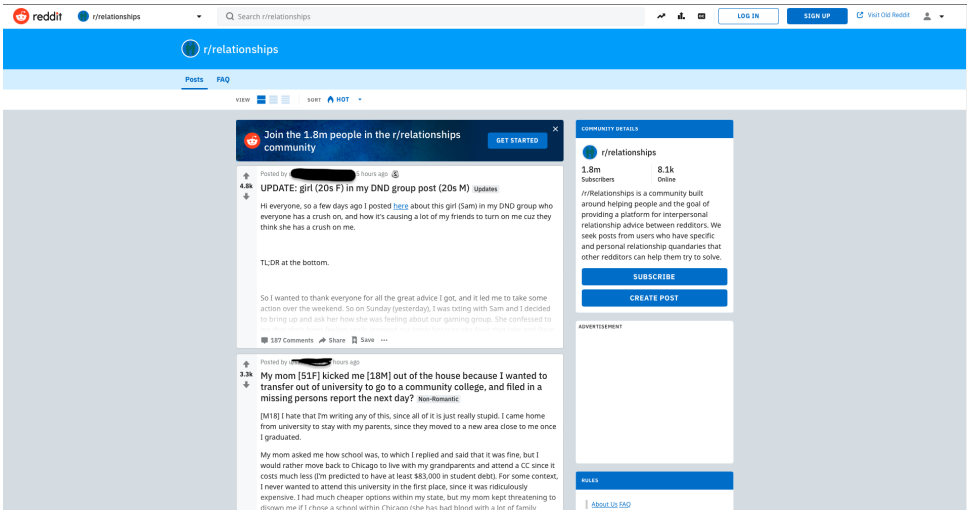

Figure 26: **Front page design of r/relationships in 2018 (Archive, 2023g).**

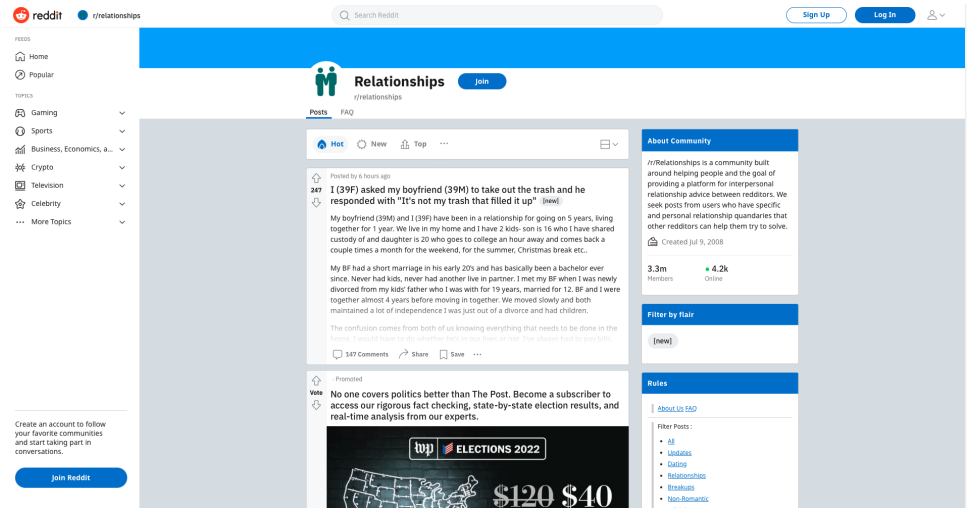

Figure 27: **Front page design of r/relationships in 2022 (Archive, 2023h).**

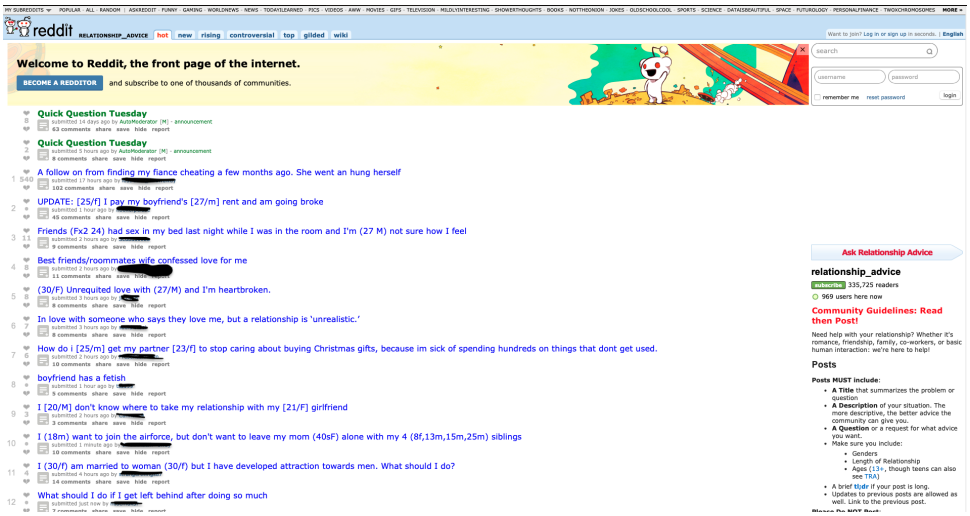

Figure 28: **Front page design of r/relationship_advice in 2017 (Archive, 2023c).**

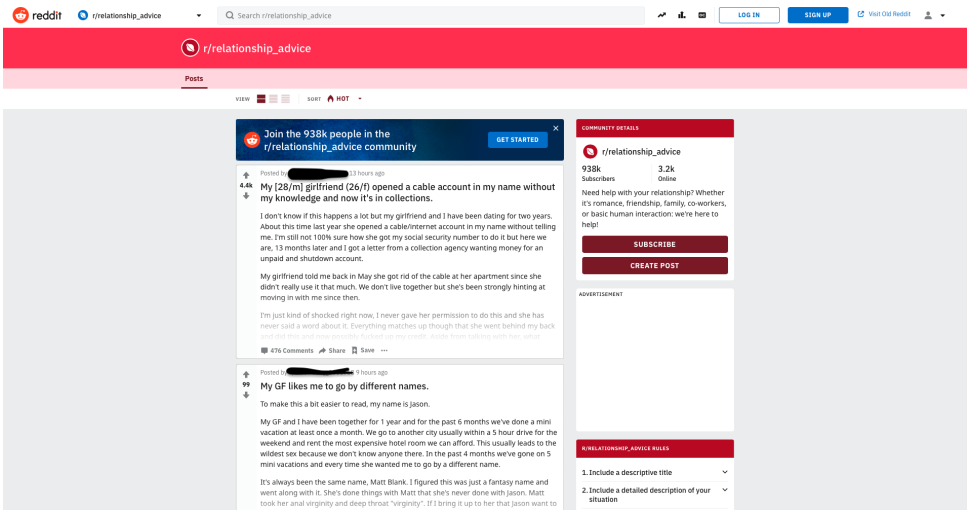

Figure 29: **Front page design of r/relationship_advice in 2018 (Archive, 2023d).**

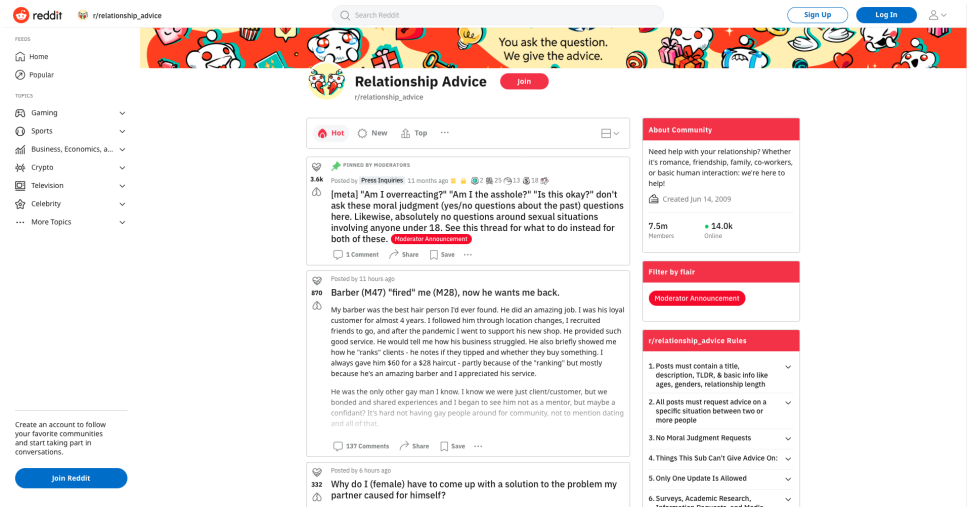

Figure 30: **Front page design of r/relationship_advice in 2022 (Archive, 2023e).**

D

D.1 Number of Comments of the Original Raw Data

Figure 12 presents the collected data before pre-processing for each subreddit separated by years.

Subreddit	Year	Total Number of Comments
r/relationships	2016	263.893
	2017	317,170
	2018	262,263
	2019	198,325
	2020	137,216
	2021	150,820
	2022	145,577
r/relationship_advice	2016	45,464
	2017	66,198
	2018	267,494
	2019	522,866
	2020	865,341
	2021	865,341
	2022	726,618

Table 12: **Total number of comments before pre-processing for the subreddit pair.**

E

E.1 Average Score of the Remaining Perspective API Attributes

Table 13 presents the remaining average toxicity attributes per year.

Subreddit	Year	Average Severe Toxicity	Average Identity Attack	Average Threat
r/relationships	2016	0.03	0.03	0.03
	2017	0.02	0.03	0.03
	2018	0.02	0.03	0.02
	2019	0.02	0.02	0.02
	2020	0.02	0.02	0.02
	2021	0.02	0.03	0.02
	2022	0.02	0.03	0.02
r/relationship_advice	2016	0.04	0.03	0.03
	2017	0.03	0.03	0.03
	2018	0.04	0.04	0.03
	2019	0.03	0.03	0.03
	2020	0.03	0.03	0.03
	2021	0.03	0.03	0.03
	2022	0.03	0.03	0.02

Table 13: **Overview of average Perspective API score for remaining toxicity attributes.**

E.2 Time Series for the Remaining Perspective API Attributes

Figures 31 to 33 represent the time series plot for the remaining toxicity attributes.

The r/relationship_advice change point is on 17.11.2019 with a *P*-value of 0.00333. For r/relationships the change points are 13.11.2018 and 15.11.2020 with the *P*-value of 0.0033 and 0.0433, respectively.

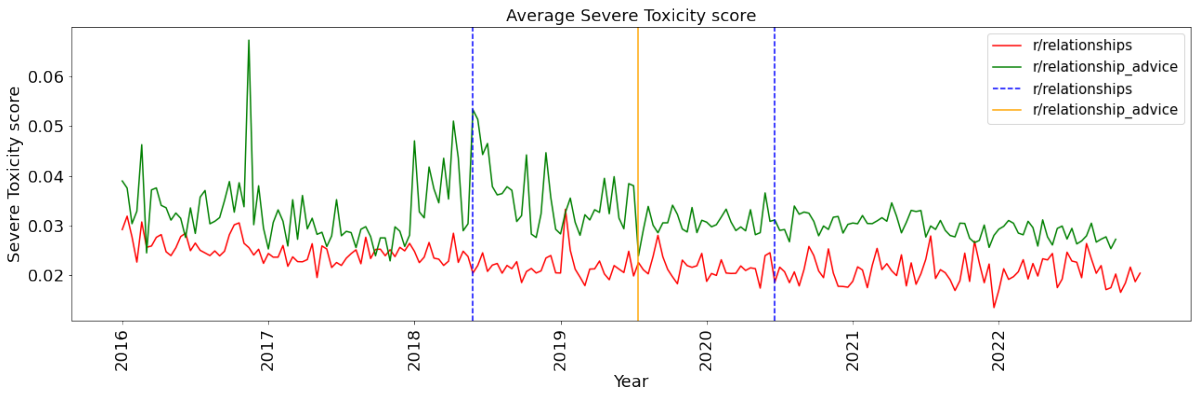

Figure 31: **Average *Severe Toxicity* score per year with change points.**

The r/relationship_advice change points are on 01.11.2018 with a *P*-value of 0.00333 and on 23.11.2020 with a *P*-value of 0.030. For r/relationships the change point is on the 22.11.2018 with a *P*-value of 0.0033.

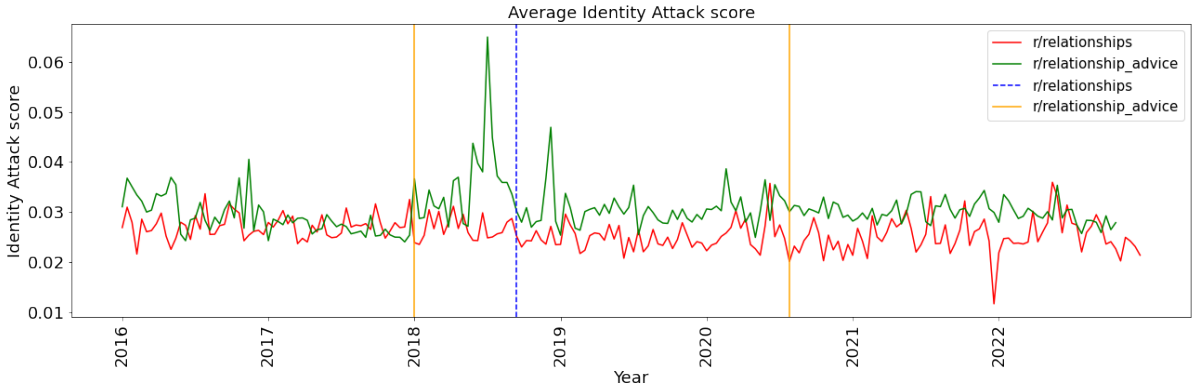

Figure 32: **Average *Identity Attack* score per year with change points.**

The r/relationship_advice change point is on 04.11.2018 with a *P*-value of 0.0033. For r/relationships the change point is on the 07.11.2018 with a *P*-value of 0.0033.

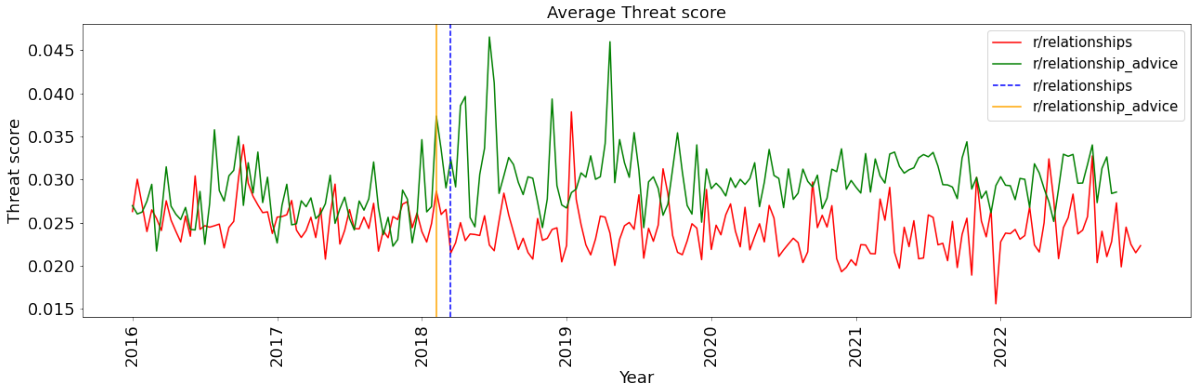

Figure 33: **Average *Threat* score per year with change points.**

E.3 Number of Toxic Comments for the Remaining Perspective API Attributes

Table 14 shows the number of comments classified as *Severe Toxicity*, *Identity Attack* and *Threat*.

Subreddit	Year	Number of Severe Toxicity Comments	Number of Identity Attack Comments	Number of Threat Comments
r/relationships	2016	16(0.007%)	11(0.005%)	9(0.004%)
	2017	4(0.002%)	7(0.003%)	15(0.006%)
	2018	8(0.004%)	5(0.002%)	11(0.005%)
	2019	8(0.005%)	5(0.003%)	12(0.007%)
	2020	7(0.006%)	6(0.005%)	5(0.005%)
	2021	4(0.003%)	3(0.002%)	5(0.004%)
	2022	7(0.006%)	2(0.002%)	7(0.006%)
r/relationship_advice	2016	4(0.009%)	1(0.002%)	8(0.019%)
	2017	16(0.026%)	8(0.013%)	11(0.018%)
	2018	95(0.039%)	63(0.026%)	79(0.032%)
	2019	139(0.031%)	85(0.019%)	82(0.018%)
	2020	123(0.026%)	57(0.012%)	98(0.021%)
	2021	185(0.025%)	98(0.014%)	135(0.019%)
	2022	101(0.018%)	60(0.011%)	84(0.015%)

Table 14: **Overview of number of comments classified as *Severe Toxicity*, *Identity Attack* and *Threat*.**

F

F.1 Number of Unique Toxic Users for the Remaining Perspective API Attributes

Table 15 illustrates the number of unique toxic users classified as *Severe Toxicity*, *Identity Attack* and *Threat*.

Subreddit	Year	Number of Unique Severe Toxicity users	Number of Unique Identity Attack Users	Number of Unique Threat Users
r/relationships	2016	1(0.003%)	2(0.006%)	0(0%)
	2017	2(0.005%)	0(0%)	2(0.005%)
	2018	4(0.010%)	2(0.005%)	4(0.010%)
	2019	2(0.006%)	0(0%)	2(0.006%)
	2020	1(0.004%)	1(0.004%)	2(0.007%)
	2021	1(0.004%)	1(0.004%)	2(0.007%)
	2022	2(0.008%)	1(0.004%)	0(0%)
r/relationship_advice	2016	0(0%)	0(0%)	0(0%)
	2017	1(0.008%)	0(0%)	1(0.008%)
	2018	32(0.053%)	18(0.030%)	31(0.051%)
	2019	33(0.037%)	21(0.024%)	18(0.020%)
	2020	29(0.033%)	9(0.010%)	17(0.020%)
	2021	29(0.024%)	14(0.012%)	25(0.021%)
	2022	22(0.023%)	9(0.010%)	15(0.016%)

Table 15: **Overview of the number of unique users classified as *Severe Toxicity*, *Identity Attack* and *Threat*.**

F.2 Average Toxicity score for the Remaining Perspective API Attributes for all Permanent Active Users

Figures 34 to 36 display the average *Severe Toxicity*, *Identity Attack* and *Threat* scores for all active permanent users.

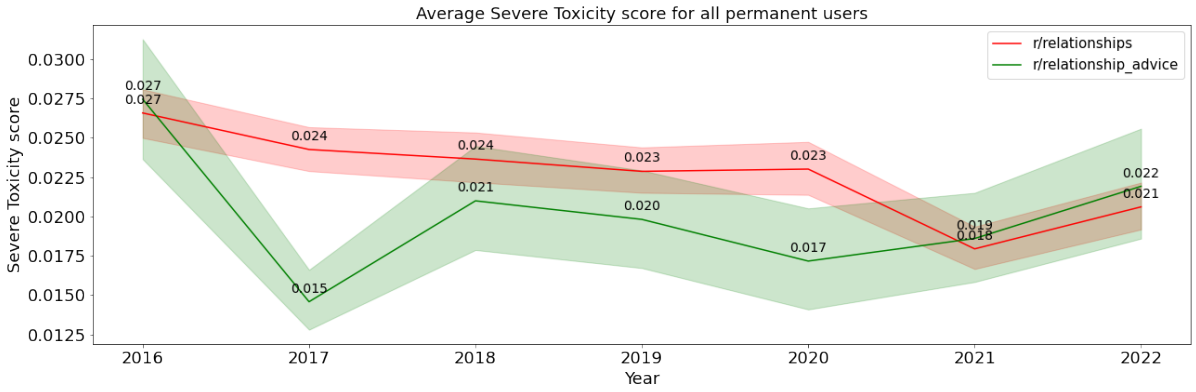

Figure 34: **Average *Severe Toxicity* score for all permanent active users.**

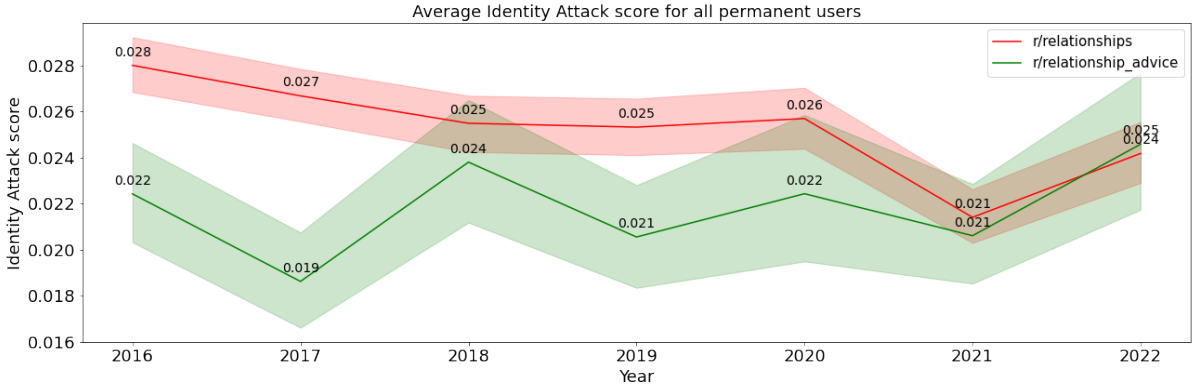

Figure 35: **Average *Identity Attack* score for all permanent active users.**

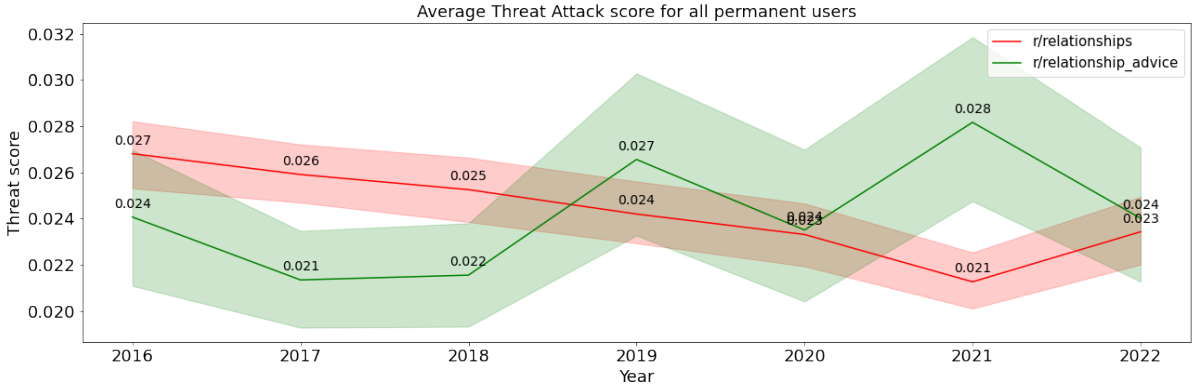

Figure 36: **Average *Threat* score for all permanent active users.**

G

G.1 Submission-Based Results for the Remaining Perspective API Attributes

Figures 16 to 46 show the submission based results for the attributes: *Severe Toxicity*, *Identity Attack Insult*, *Profanity*, and *Threat*.

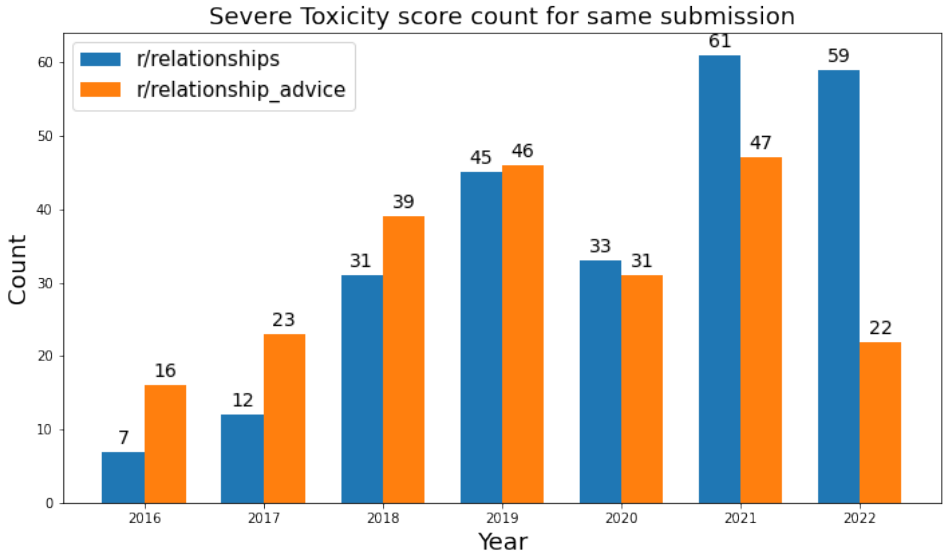

Figure 37: ***Severe Toxicity* counts for submission-based analysis.**

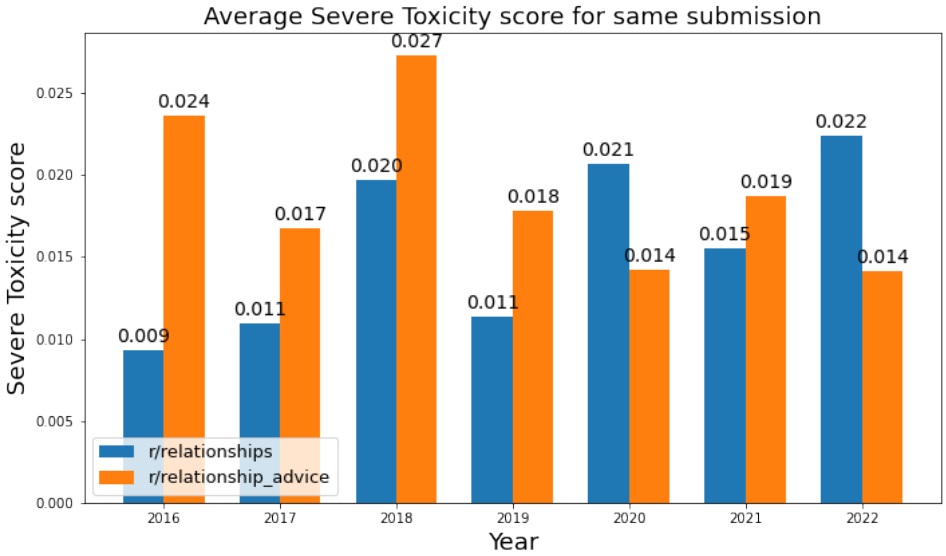

Figure 38: **Average *Severe Toxicity* score for submission-based analysis.**

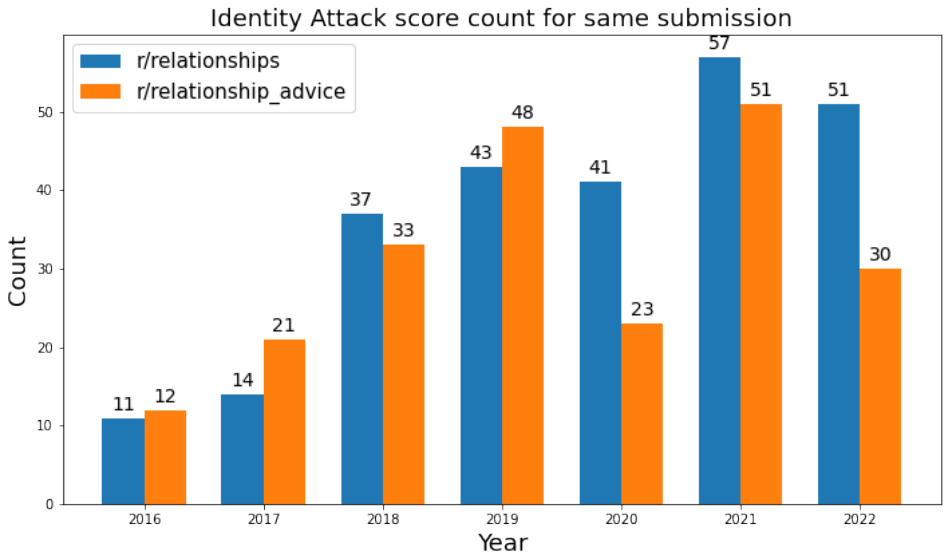

Figure 39: ***Identity Attack*** **counts for submission-based analysis.**

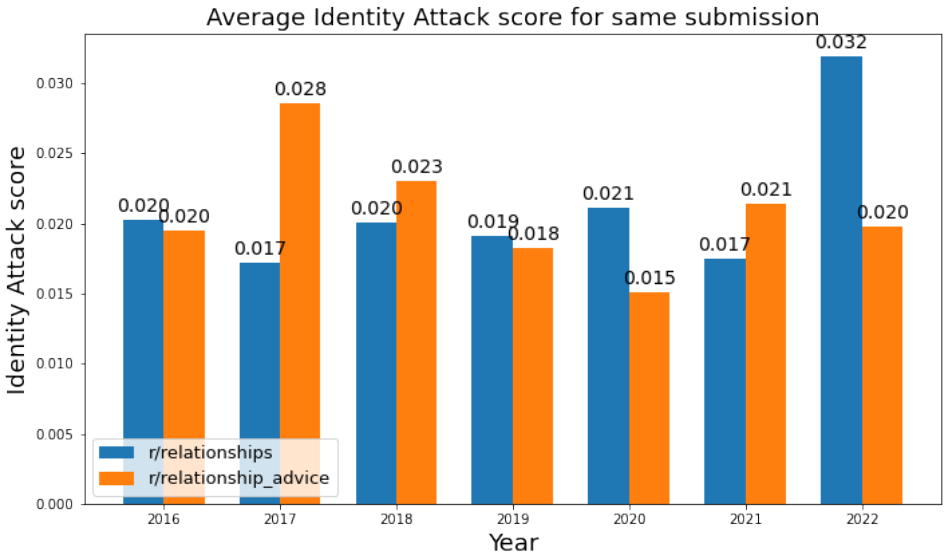

Figure 40: **Average *Identity Attack* score for submission-based analysis.**

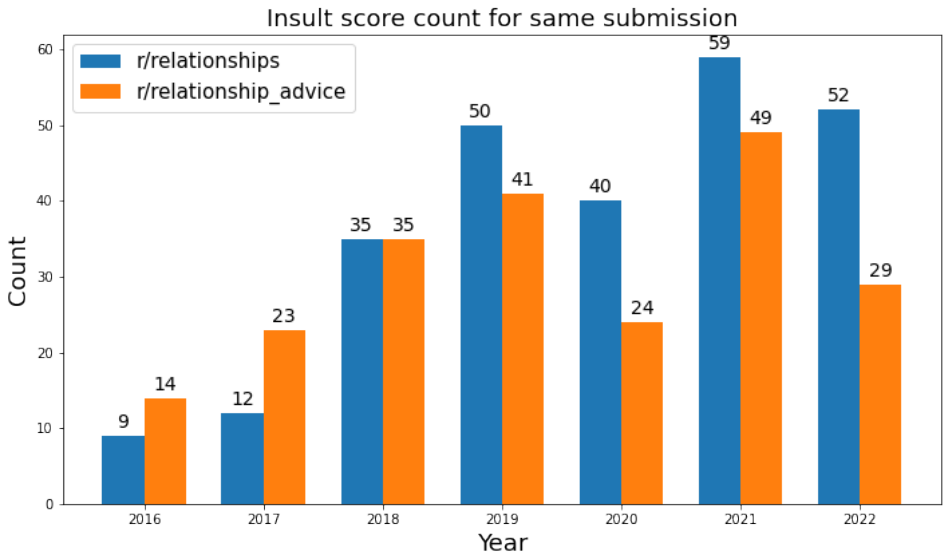

Figure 41: ***Insult* counts for submission-based analysis.**

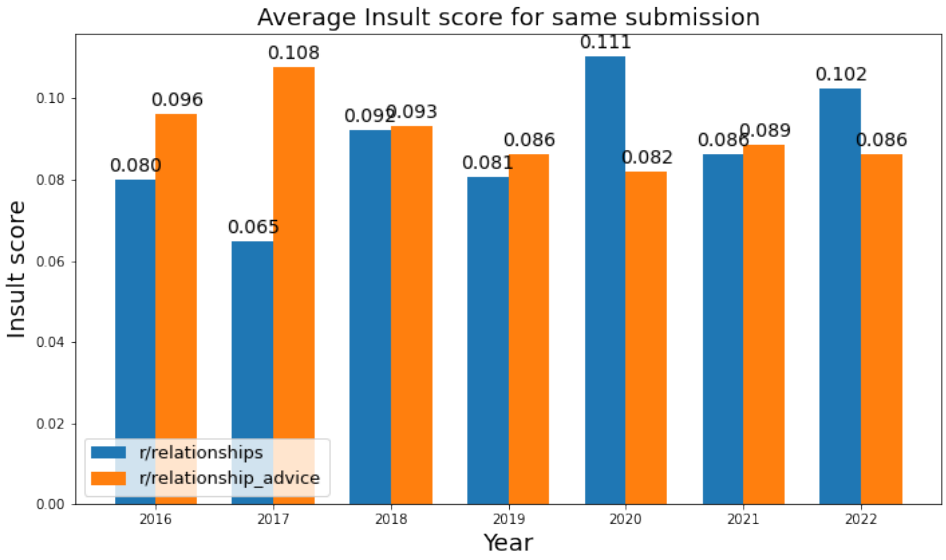

Figure 42: **Average *Insult* score for submission-based analysis.**

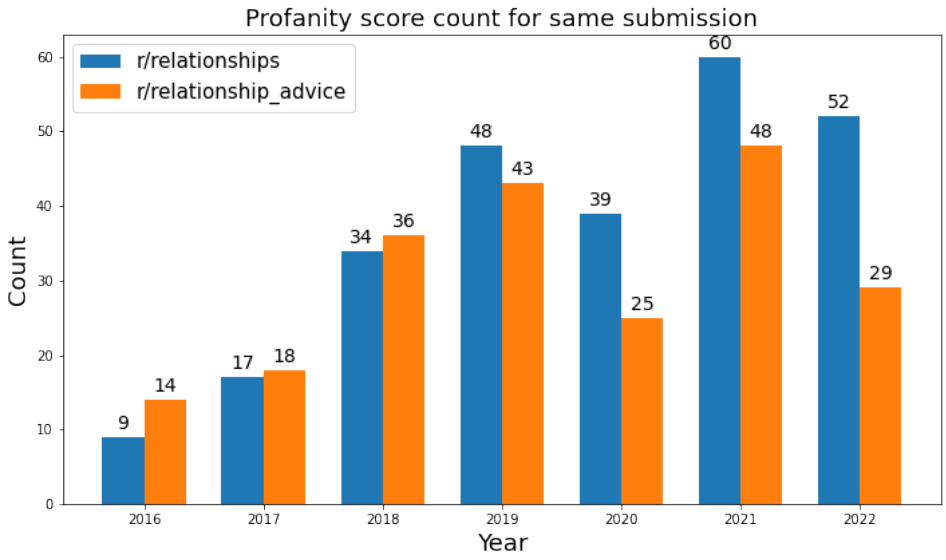

Figure 43: ***Profanity* counts for submission-based analysis.**

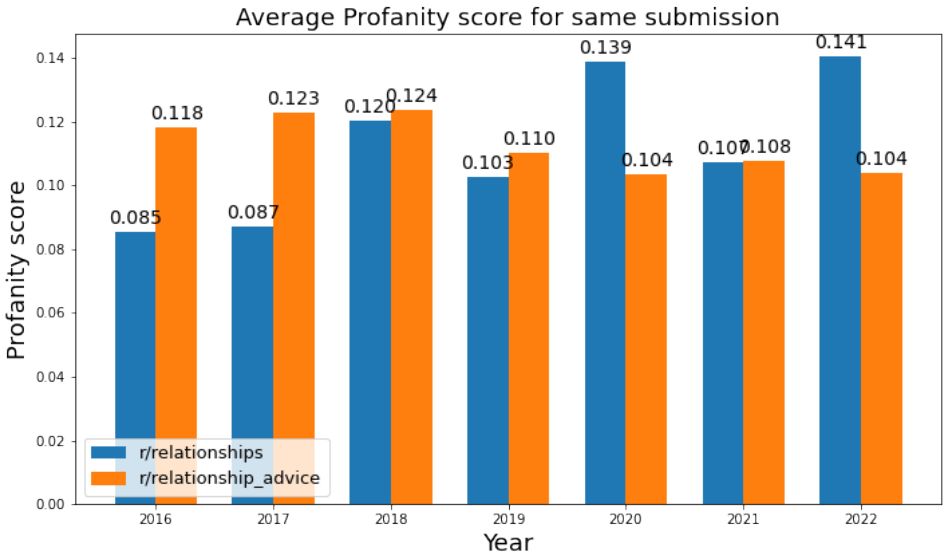

Figure 44: **Average *Profanity* score for submission-based analysis.**

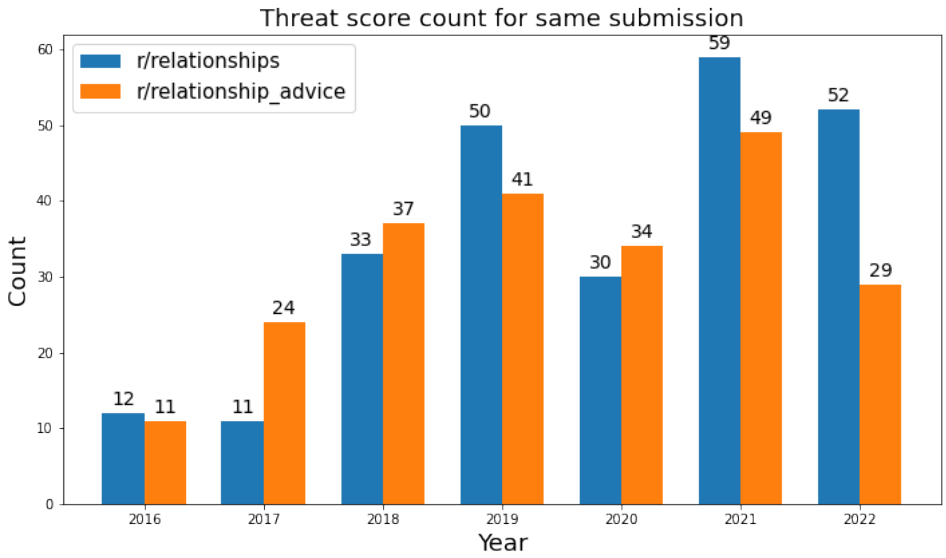

Figure 45: ***Threat* counts for submission-based analysis.**

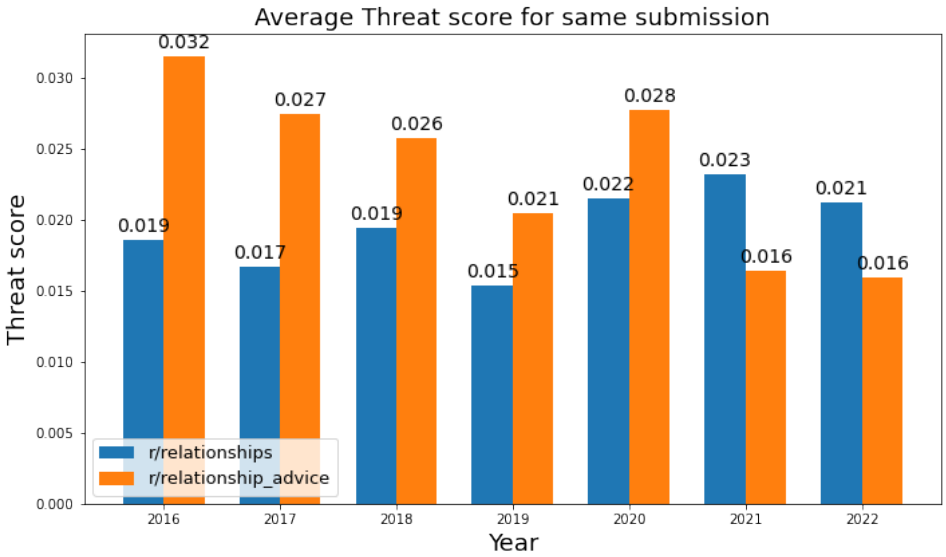

Figure 46: **Average *Threat* score for submission-based analysis.**

H

H.1 Total Number of Deleted Users and Removed Comments

Table 16 displays the total number of deleted users and the total number of removed comments (after pre-processing).

Subreddit	Year	Total Number of Removed Comments	Total Number of Deleted User
r/relationships	2016	16k	3k
	2017	32k	5k
	2018	24k	3k
	2019	11k	5k
	2020	7k	3k
	2021	10k	3k
	2022	9k	2k
r/relationship_advice	2016	50	2k
	2017	315	1k
	2018	3k	6k
	2019	6k	19k
	2020	6k	24k
	2021	19k	27k
	2022	61k	11k

Table 16: **Overview of deleted user and removed comments (after pre-processing) for each subreddit.**

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Zuhilfenahme anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum Unterschrift