CHARLES UNIVERSITY

Universität Konstanz

Master Thesis:

From Supreme Gentlemen to Incel Rebellion – Analysing the Radicalisation Potential of the Incel Community on Twitter

written by Mia Nahrgang

Charles University

Department: Security Studies

Supervisor: doc. PhDr. Vít Střítecký

Student-ID: 40955907

University of Konstanz

Department: Politics and Public Administration

Supervisor: Prof. Nils B. Weidmann

Student-ID: 1066432

Date: 10.07.2022

Contents

List of Figures and Tables

Abstract:

Elon Musk in April 2022 surprisingly declared his intention to buy Twitter with the goal to ensure free speech. However, maybe ensuring free speech ensuring free speech on a powerful social media platform with 229 million active users per day is not a risk-free endeavour. The focus of this thesis is the incel community, which revolves around shared frustrations about failing to achieve sexual relations, opposition to feminism and violence-inciting misogynism. I ask the question: To what extent do more radical tweets diffuse further within the incel community? More concretely, I quantitatively investigate the relationship between the toxicity and the misogynism of a tweet and the number of times it is retweeted on a self-collected dataset encompassing 52,927 tweets. My findings suggest that toxic and misogynist tweets are retweeted more often and thus do spread further within the incel community on Twitter. This has crucial implications for the radicalisation potential of the incel community on Twitter as frequent exposal to radical content might amplify the radicalisation of others.

Keywords: Incels – Information diffusion – Radicalisation – Echo-chambers – Twitter

1. Introduction

> You go on Twitter, and you see someone post something interesting and it barely gets any retweets or likes. If you post something inflammatory...it rapidly gets retweeted or liked.
> (Incel Interviewee cited in Daly & Reed, 2022, p. 26).

> Police Recording of Alex Minassian immediately after his attack in 2018. He answers the question of what he thought when he learned about Rodger Elliot's incel-motivated attack in 2014:
>
> Alek Minassian: I felt kind of ah proud of him for ah his acts of bravery.
>
> Detective: Okay alright and what about ah how you started to, to, to change your thinking. Was, was any was, was any of that going on?
>
> Minassian: I was starting to feel ah radicalized at that time.
>
> Detective: You were eh okay and when you say radicalized what do you mean by that?
>
> Minassian: Meaning I felt it was time to take action and not just sit on the side lines and just ah fester in my own sadness.
>
> (A. Minassian, personal communication, 23 April 2018, p. 116).

In April 2022 Elon Musk surprisingly declared his intentions to buy Twitter, a short message social media provider and microblogging service founded in 2006 on which public posts are called tweets. At the time of the submission of this thesis, it is not yet foreseeable whether he will finally go through with his plan. Musk's primary intention for the take-over Twitter is to ensure free speech and to limit the companies' moderation (Kleinman, 2022) as the platform in the past has banned controversial accounts including former US President Donald Trump, the conspiracy theorist Alex Jones, or the prominent right-wing figure Milo Yiannopoulos (Milmo, 2022) as well as accounts managed by the Islamic State (BBC, 2016). However, it remains questionable whether ensuring free speech on a powerful social media platform with 229 million active users per day[1] is a risk-free endeavour and does not bring potential dangers. The European Union chose a more critical approach toward free speech on the internet and agreed in the same month on its Digital Services Act which aims at limiting the spread of illegal content online so that eventually a safer digital space is created in which the fundamental rights of all users are protected. This new legislation affects companies with more than 45 million monthly active users and thus also addresses Twitter.[2] The combination of these two events shows that it is crucially relevant to investigate critical online content. This thesis aims to do so by focusing on the incel community.

The incel community is part of the so-called manosphere, which is a technological phenomenon that consists of multiple subgroups and goes back to the 1960s and 1970s. According to Hermansson et al. (2020, p. 163), the manosphere refers to a "loose collection of websites, forums, blogs, and vlogs concerned with men's issues and masculinity, oriented around an opposition to feminism". At the time of its origin the manosphere, as it essentially revolves around men's rights, had links to second-wave feminism. However over time, the focus shifted and subsequently women and female empowerment were identified as the origin and source of men's problems (Horta Ribeiro, Blackburn, et al., 2021). The manosphere consists of multiple sub-communities, unified by common key features - the feeling that masculinity is threatened by women and an aversion towards feminism which is described as hypocritical and oppressive (ibid.). One older subcommunity of the manosphere is the 'Pick-Up Artists' (PUA) community which centres around influencing women into having sex while disregarding the concept of consent (Hermansson et al., 2020). The community is all about the

[1] Statista. Number of monetizable daily active Twitter users worldwide from 1st quarter 2017 to 1st quarter 2022. Retrieved from: https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/, checked 31.05.2022.

[2] European Commission. The Digital Services Act package. Retrieved from: https://digital-strategy.ec.europa.eu/-en/policies/digital-services-act-package, checked 04.05.2022.

‘game’ to pick up women. Another subcommunity consists of the ‘Men’s Rights Activists’ (MRAs) which believe that: men are “undergoing systematic oppression at the hands of feminism and progressive movements” (ibid., p. 165). Additionally, ‘Men Going Their Own Way’ (MGTOW) is a sub-community of the manosphere, that is characterised by their attempt to minimise contact with women in general because they believe that society is rigged against men (Horta Ribeiro, Blackburn, et al., 2021). The last subcommunity and the focus of this thesis is the incel community. Incel is an abbreviation for involuntary celibate and the community goes back to a website created by a young female student in 1997 with the aim to offer a platform for expressing frustrations and support for individuals failing to achieve sexual relations. However, over time parts of the incel community underwent a radical militarisation, advocating misogynism and open violence against women and attractive men (Ging, 2019). The ideological base of incels is the firm conviction of a biologically determined hierarchy centred on attractiveness, where idealised men and women, so-called *Chads* and *Stacies* are positioned at the top, followed by *normies* or *betas*, leaving the bottom for incels. Consequentially, most of the women are attracted by *Chads*, leaving only the judged less attractive ones to *normies* and none for incels. Crucial is that according to the incel perception women are responsible for their loneliness and rejection as they follow their barbaric instincts and fail to appreciate the *supreme gentlemen* (Hoffman et al., 2020, p. 567). Moreover, incels believe that in the old traditional society, they would have been romantically successful. However, feminism and a progressive society in which women can freely chose whom they wish to partner with, incels remain leftover (Rouda, 2020). A common theme within the manosphere that arose out of the incel ideology is the pill analogy, which has its origin in the *blue pill/red pill* dichotomy from the movie the Matrix. Taking the blue pill means living in an illusion, whereas taking the red pill symbolises an understanding and awakening. In the context of the manosphere taking the red pill symbolises a revelation of the truth about women and society (Hoffman et al., 2020). Additionally, there is a new creation from the manosphere: taking the *black pill*, which means ultimately accepting the “reality where women and society are intrinsically biased against men who lack specific physical attributes, who therefore have no hope of ever being attractive to women or even accepted by society” (ibid., p. 568).

That online incel radicalisation is a serious threat in that it can have dangerous real-life implications has been underlined by three clear incel-motivated terrorist attacks. The first terrorist attack by a self-declared incel happened in 2014 in Isla Vista, California where Elliot Rodger killed six people and wounded 14 but ultimately failed to enter a sorority house, which he intentionally selected as it was known to have the “most beautiful girls” as members (Rodger, 2014, p. 132). His attack was driven by revenge against women who had rejected him his whole life (Hermansson et al., 2020). This conclusion can be drawn as Rodger left behind extensive documentation of his thoughts. Immediately prior to his attack, he posted a video on YouTube that he called "Day of Retribution". Moreover, he sent via mail a 137-pages manifesto with the title "My Twisted World" to his immediate social environment (Allely & Faccini, 2017). In this manifesto, he describes himself as an "ideal, magnificent gentleman" (Rodger, 2014, p. 109) and describes his desire to have a "beautiful, tall blond-haired girl" as his girlfriend (ibid., p. 76). Additionally, he describes his radicalisation process and concludes that "if I'm unable to have such a life, then I will have no choice but to exact revenge on the society that denied it to me" (ibid., p. 109). Ultimately, Rodger aims at containing the wickedness of women (White, 2017) as otherwise "the whole of humanity will be held back from a more advanced state of civilization". Rodger soon became idealised as a "patron saint" and ever since "goingER" (for Elliot Rodger) is an expression used in the incel community to describe following Rodger's example to commit incel martyrdom (Baele et al., 2021). In April 2018 Alek Minassian then killed ten people and injured 16 more in Toronto, Canada with a rented truck. In the same year in November Scott Beierle killed two women at a Hot Yoga studio in Tallahassee, Florida (Hoffman et al., 2020). All three of them were not only part of the incel community but also actively interacted on online platforms. Moreover, Baele et al. (2021, p. 1667) claim that although they eventually attacked alone, they were "part of a broader community of like-minded individuals with a vivid online presence". Elliot Rodger and Scott Beierle posted openly misogynist YouTube videos before committing their attacks and Alek Minassian referred in his final Facebook post to Rodger as *Supreme Gentlemen* and declared that the incel rebellion had begun (Hoffman et al., 2020). Moreover, Minassian also was in contact with Rodger on Reddit, with whom he exchanged frustrations (A. Minassian, personal communication, 23 April 2018). More crucially Minassian described in his interrogation how witnessing Rodger's radicalisation and eventually his attack crucially accelerated his own radicalisation process, leading him to the conclusion that he himself needed to act upon his misery (A. Minassian, personal communication, 23 April 2018). This is an interesting finding as it suggests that the diffusion of incel content online has severe implications for the future as it might amplify the radicalisation of others. This thesis thus addresses the diffusion of incel related content in combination with online radicalisation.

Although there is a lot of research on the incel community on various platforms like Reddit (Farrell et al., 2019; Massanari, 2017), YouTube (Papadamou et al., 2021) or incel-specific Forums (Baele et al., 2021; Horta Ribeiro, Blackburn, et al., 2021; Jaki et al., 2019) and a strong scholarly focus on the automated detection of incel hate speech on Twitter (Frenda et al., 2019; Hajarian & Khanbabaloo, 2021; Jaki et al., 2019; Sang & Stanton, 2020), there remains a research gap about the radicalisation potential of the incel community on Twitter. Researching this gap is however important for two reasons. First, Twitter has compared to other incel-specific forums a broader audience and incel-related content can thus be accessed with lower barriers. The importance of Twitter should therefore not be underestimated as the social media platform could portray an important step in the initial radicalisation. Secondly, the display of radical misogynist content on Twitter, a public and open accessible platform, would portray a further development toward becoming a mainstream ideology. It is thus important to assess the radicalisation potential of the incel community on Twitter. Additionally, on a more practical note, information diffusion, which is the focus of this thesis, can easily be measured on Twitter by focusing on the number of retweets a tweet receives. Twitter itself defines a retweet as "[a] Tweet that you share publicly with your followers" and at the same time states that retweeting is "a great way to pass along news and interesting discoveries on Twitter".[3] However, the retweet function has been the centre of some controversies. The creator of the retweet button himself, Chris Wetherell, states that he regrets having invented this function. He argues that although it is an effective tool for information diffusion it caused some serious side effects in that it is known to "incentivize extreme, polarizing, and outrage-inducing content" (Kantrowitz, 2019). This is the case as the retweet button encourages sharing without taking time to consider all aspects. The creator thus concludes that Twitter since the implementation of the retweet function has become an "anger video game" where "retweets were the points" (ibid.). This description fits well with the literature on information diffusion which will be the theoretical basis of this thesis and overall agrees that negative, hateful, or angry content is retweeted more often. Accordingly, this thesis asks the question: *To what extent do more radical Tweets diffuse further within the incel community?* This question will be analysed by testing two hypotheses claiming that facilitated through an echo chamber effect more toxic and more misogynist tweets are retweeted more often within the incel community on Twitter.

These hypotheses are tested through a self-collected dataset encompassing 52,927 tweets from 11,637 users in the period of seven weeks from March 21, 2022, to May 06, 2022. The tweets collection was based on 12 keywords which are used frequently within the incel community on Twitter. The toxicity of a tweet is consequentially determined by the severe toxicity attribute of the Perspective API (*Perspective API*, 2018) which is a machine learning

[3] Twitter. How to retweet. Retrieved from: https://help.twitter.com/en/using-twitter/how-to-retweet, checked 08.06.2022.

algorithm. The misogynism of a tweet on the other side is determined with a dictionary approach based on a dictionary measuring misogynism developed by Farrell et al. (2019).

This thesis is structured as follows. First, the author is going to review the existing literature thereby especially focusing on digital violence against women, the incel community and literature theorising Twitter as well as extremism and radicalisation processes. Afterwards, the theoretical considerations, namely literature on information or tweet diffusion and the echo-chamber effect are discussed before the hypotheses are formulated. Chapter four will focus on methodological considerations including the data collection process, the definitions and conceptualisation of the variables, and the methodological tools employed to test the hypotheses. Chapter five presents the results which are subsequentially discussed, while attention is paid to limitations and future research ideas. Finally, the conclusion summarises the findings and provides an outlook.

This thesis found that toxic and misogynist tweets are retweeted more often within the incel community on Twitter. More concretely, an increase in the toxicity score of a tweet is associated with an 18.2% increase in retweets. Similarly, misogynist tweets are retweeted 3.2% more often than non-misogynist tweets. This has crucial implications on the radicalisation potential as frequent exposal of radical incel content can amplify the radicalisation of others.

2. Literature Review

This thesis is based on the scholarly debate, which will be reviewed in the following chapter. The author is going to touch on the literature about hate speech and digital violence against women as well as the literature about the manosphere and incels more specifically. Then, the author is going to theorise about Twitter before reviewing literature about extremism, the phenomenon of radicalisation and lone-wolf terrorism. Ultimately, existing gaps in the literature will be highlighted.

2.1 Hate Speech and Digital Violence against Women

When approaching the topic of this thesis it is imperative to consider the literature on hate speech and digital violence against women. Silva et al. (2016, p. 688) provide a general definition of hate speech and define it as “any offense motivated, in whole or in a part, by the offender’s bias against an aspect of a group of people”. Moreover, they investigate the targets of hate speech on Twitter and Whisper, an anonymous social media site. They find that hate speech caused by race is the most common type on Twitter followed by behaviour, physical, sexual orientation, class, ethnicity, and gender is only in seventh place, which is then followed by disability, religion and other. Regarding Whisper, the ranks remain similar. However, gender is only in eighth place.

While hate speech motivated by gender seems not to be as widespread as other types of hate speech it does have very specific characteristics. Mantilla (2013, p. 564) argues that what she calls “gendertrolling” is fundamentally different from regular trolling as it is “dramatically more destructive to its victims”. Moreover, she identifies six characteristics of “gendertrolling”. According to her, it is exercised in a coordinated manner by multiple participants, sometimes dozens or even hundreds of trolls. The sheer numbers then allow the trolls to overwhelm and flood the victim with attacks. Secondly, insults are gender-based and include pejorative terms like *slut* or *whore*. Thereby special attention is directed to degrading the woman’s physical appearance. Additionally, “gendertrolling” is characterised by its usage of vicious language and extensive threat descriptions – be it sexually or physically. Also, threats are often not only vicious but also credible and underlined by revealing personal information like the home or work address of a victim. Furthermore, “gendertrolling” is specific in that the attacks last unusually long, sometimes multiple years, and are carried out on multiple platforms and even in the offline world. Finally, “gendertrolling” is different in regard to the trigger of hate speech which is almost exclusively women speaking out about some sort of sexism or advocating for feminism. Therefore Mantilla (2013, p. 569) concludes that “gendertrolling” is “something above and beyond generic online trolling and a phenomenon that, not dissimilar to street and sexual harassment, systematically targets women to prevent them from fully occupying public spaces”. The explicit target of “gendertrolling” is thus to keep the online world a male-dominated sphere (ibid.).

Similarly, Jane (2014) describes recurring characteristics of what she calls “e-bile” where anonymous attackers target women who are standing in the public. An essential part of the attack is the usage of “sexually explicit rhetoric” (Jane, 2014, p. 560) and comments regarding the physical appearance of the victim. Additionally, the targets are “hypersexualised as ‘sluts’ and then derogated for being ‘sluts’ who did not pass muster because they were too ugly, too fat, too small breasted, too old, too lesbian and so on” (ibid.). Jane, furthermore, argues that the “e-bile” generates a sort of competition between attackers on who can make the most offensive insults and thus includes the potential to escalate. Moreover, Jane contends that “e-bile” has transformed. She states that before 2011 attacks were directed at a relatively narrow set of victims. This, however, changed after 2011 and since then “e-bile” is directed at all sorts of women. At the same time, “e-bile” became more mainstream, widespread on public platforms and toxic. Thus, Jane concludes that: “gendered e-bile has now become normalised such that it is now acceptable to express even the most minor disagreement through the most affronting, offensive and aggressive sexualised venom” (ibid., p. 566).

Poland (2016, p. 3) directs her book at what she describes as cybersexism which is “the expression of prejudice, privilege, and power in online spaces and through technology as a medium” and more specifically she focuses on “verbal and graphic expression of sexism in the form of online harassment and abuse aimed at women” (ibid.). She argues that to understand what drives cybersexism, it is necessary to fully understand sexism in its offline version. Moreover, she claims that both are driven by the desire to underline male dominance. Consequentially, “activities aimed at building and reinforcing male dominance online are conducted in order to re- create the patterns of male domination that exist offline” (ibid., 5). Additionally, Poland provides a classification of different types of cybersexism starting from relatively harmless practices like mansplaining and derailing with the latter being a tactic of interrupting and refocusing a conversation to display the expertise of the interrupter. Other types with more crucial implications are gendered abuse and harassment, online threats, including threats of death and rape, and finally forms of cybersexism which result in direct real-life consequences. This could be doxxing, which is the revelation of sensitive private data such as the full name, home address, social security number, credit card details, or SWATing which refers to making emergency calls for the released home address, so that a police unit is sent to the respective address.

Powell and Henry (2017, p. 5) on the other hand focus on a more extreme form of digital violence against women, namely 'technology-facilitated sexual violence', which they define as "a concept […] to refer to the diverse ways in which criminal, civil or otherwise harmful sexually aggressive and harassing behaviours are being perpetrated with the aid or use of digital communication technologies". Additionally, they claim that the distinction between online and offline forms of violence or harassment is increasingly outdated as "digital technologies play an increasingly central part in where and how we work, learn, play and communicate" (ibid., p. 51). Moreover, "[d]igital technologies provide sites or mechanisms for the construction of our identities and relationships, as well as of our professional and social lives" (ibid.). Therefore, it is important to understand that the consequences of online harassment are not limited to the online realm but instead also transform into offline lives.

Furthermore, there is literature focusing on specific harassment campaigns. For example, Aghazadeh et al. (2018) present a case study of the prominent harassment campaign #GamerGate from 2014 targeting female and minority game developers and their allies, which has been associated with the manosphere. The harassment campaign started when Zoe Quinn's ex-boyfriend accused her publicly of having an affair with a video game journalist in exchange for positive reviews for her newly released video game Depression Quest. The ex-boyfriend linked the post with his accusations to 4Chan where like-minded users gathered and started their attack on Quinn. The attack became especially serious when Quinn's personal information including her address and social security number was revealed. With growing strength, criticism of the movement grew as well, and users and gaming journalists started to protest the open display of misogynism. However, the support for Quinn backfired and allies became themselves targets of online harassment. The attack was given its name on Twitter with #GamerGate by the actor Adam Baldwin who, with the help of his popularity, pushed the movement to a new scope and opened it to the far-right. Apart from 4Chan, Reddit, 8Chan and YouTube, Twitter was also one of the platforms used to spread hate. Aghazadeh et al. (2018, p. 187) conclude that as a long-term consequence of #GamerGate "online harassment to silence minorities and especially women is no longer an odd episode, as was initially the case with GamerGate, but has become a *normal* part of the Internet."

Massanari (2017, p. 333) furthermore describes the events surrounding #GamerGate as part of a "toxic technoculture" that "relies on Othering of those perceived as outside the culture, reliance on outmoded and poorly understood applications of evolutionary psychology, and a valorization of masculinity masquerading as a peculiar form of 'rationality'". Apart from #GamerGate she also focused on the event called, the Fappening, where illegally acquired nudes from celebrities were circulated and hatefully discussed on 4Chan and Reddit. She investigates through a long-term participant observation how Reddit's design and algorithm created a hub for anti-feminist and misogynistic activities. This is the case as Reddit's karma point system affects the display of subreddits on the front page and highly upvoted posts tend to receive more attention. According to Minassian, this combined with the ease of creating an anonymous account and loose content moderation regulations helps to flourish the "toxic technoculture" on Reddit.

Additionally, Hardaker and McGlashan (2016) investigate through a combination of computational linguistics and discourse analysis the phenomenon of rape threats on Twitter on the example of a harassment campaign against feminist journalist Caroline Criado-Perez in 2013. She advocated for an additional woman on the English banknote to strengthen gender representation in the British currency and as a result became the target of a harassment campaign on Twitter, which included threats of rape and murder, as many as fifty threats an hour, and even involved bomb threats. The authors collected about 76,000 tweets over three months and focused on sexually aggressive online behaviour and with what the terms abuse, rape, threat, and trolls co-occurred. They find that sexually aggressive behaviour is used as a misogynistic weapon to control the discourse of women online.

Moreover, Bartlett et al. (2014) aimed at revealing the volume, degree and type of misogynist language employed on Twitter by analysing 108,044 tweets posted that included misogynist/ sexist terms of users based in the UK. Among their findings is that from all the tweets including the word *rape* approximately 12% appear to be threatening. Additionally, 18% of tweets including the terms *whore* or *slut* are classified as misogynist. Finally, they also find that increases in the usage of sexist language can be driven by media reporting about events related to sexism.

Filippo et al. (2015) finally demonstrate that online hate speech can transform into dangerous real-life consequences by investigating the association between misogynism on Twitter and rape statistics in US federal states. When removing Washington DC as an outlier they find a significant association of a Pearson correlation of $r = 0.36$ ($p < 0.01$) and thus, conclude that social media can be used as a predictor of criminal behaviour.

2.2 The Manosphere and the Incel Community

The second part of the review is going to focus on previous work about the manosphere and specifically the incel community.

Ging (2019) conducts a qualitative thematic analysis of various web pages associated with the manosphere and identifies key categories and features of the same while characterising the portrayed masculinities within this online space. She classifies five categories of interest groups within the manosphere and identifies them as MRAs (men rights activists), MGTOW (men going their own way), PUAs (pick-up artists), traditional Christian conservatives and gamer/geek culture. The last category encompasses the incel community and Ging (2019, p. 651) describes members of this community as "hybrid masculinities, whose self-positioning as victims of feminism and political correctness enables them to strategically distance themselves from hegemonic masculinity, while simultaneously compounding existing hierarchies of power and inequality online". She finds that social media has facilitated the spread of antifeminist ideas and information across communities, platforms, and geographical borders. Moreover, she argues that the manosphere is increasingly characterised by extreme misogyny and proneness to personal attacks.

Daly and Reed (2022) conduct interviews with ten self-identified incels which they choose through their Twitter activity. They then applied the hegemonic masculinity framework (Connell & Messerschmidt, 2005), which assumes the existence of multiple masculinities which are organised by a social hierarchy where the hegemonic masculinity dominates other forms of masculinities (subordinated or marginalised) as well as queer persons and women. Daly and Reed find the following five themes that summarise the incel experience: masculinity challenges, subhuman status and social rejection, the BlackPill, shit-posting, and perceived effects of inceldom. Accordingly, incels feel challenged in their masculinity due to lacking sexual experience as well as deficient physical appearance or mental health. Consequentially, the incel experience is characterised by feeling treated as subhuman which is a result of the challenged masculinity. An additional theme is the BlackPill which is constituted by "a three-step process: it is not only the belief that looks matter most, but also the use of scientific evidence (and individual experiences) to support the ideology which ultimately leads to internalization and acceptance of their fate" (ibid., p. 23). The most significant finding of Daly and Reed is that the process of "shit-posting" which they describe as the "[u]se of violent, misogynistic, racist, or generally unacceptable rhetoric on forums or social media" (ibid., p. 20), which is intentionally used to shock and provoke people outside the inceldom and allows incels "to generate a localized, dominant masculinity online" (ibid., p. 26) although they feel marginalised most of the time. This is especially true for public forums like Twitter where "shit-posting" is used by incels to prove their masculinity. The last theme identified by Daly and Reed is the perceived effects of the inceldom, namely depression, sadness, a feeling of isolation and a tendency to suicidal thinking.

Speckhard et al. (2021) in contrast conduct a quantitative survey with 272 self-identified incels as participants, which they reached by distributing the survey on an incel forum. Apart from socio-economic demographics, they ask about incel ideology, attitudes towards violence and psychological symptoms. Their research aims at establishing whether the broader incel subculture represents a threat to society and potentially embodies the characteristics of a terrorist movement as some high-profile cases and the most extreme fringes let assume. Their findings confirm that the incel community does revolve around the blackpill notion. Moreover, their results show that almost the majority (46.3%) of participants completely disagree with the statement that incels are "willing to endorse violence.". However, a crucial share of 17% of the respondents does agree with this statement. At the same time, 26.1% of the questioned people agree with the statement "I sometimes entertain thoughts of violence toward others" and 13.6% agreed to some extent with the statement, "I would rape if I could get away with it". Moreover, 31 participants claimed to admire Rodger Elliot for his attack. It is interesting that at the same time 82% claim to completely disagree with the Canadian decision to label incels as a terrorist group. The most crucial finding of Speckhard et al. (2021) is that participants who do understand themselves as dangerous claim that the forum made them feel more violent. The same is true for participants who self-assess as highly misogynist. These participants as well claim that the forum made them feel even more misogynist as the forum tends to reaffirm and validate their views.

Furthermore, O'Malley et al. (2022) employ an inductive qualitative analysis of over 8,000 posts from two online incel forums and analyse the norms, values, and beliefs of incels from a subcultural perspective. They identify the following five topics: the sexual market, women as naturally evil, legitimising masculinity, male oppression, and violence whereas the first four are used to validate and justify the last one. The idea of the sexual market revolves around the assumption that it is female-led in a way that women have the privilege to choose partners for sexual relations and thus act as "sexual gatekeeper by deciding who they reject and with whom they have intercourse" (ibid., p. 10). Additionally, women are depicted as naturally evil as they are driven "by the desire for reproduction and are narcissistic and cruel in pursuit of these goals" (ibid., p. 13). Furthermore, masculinity is legitimised by highlighting the intellectual inferiority of women. Despite being legitimate, incels experience a feeling of oppression by both hegemonic and biological superior men and modern feminist women. The combination of the four experiences thus ultimately leads to the legitimisation of violence and a desire for revenge. Additionally, they focus on the role of the internet in enhancing extremism and radicalisation online as it facilitates the formation of subcultures where beliefs are shared and endorsed, and a commonly agreed value system might justify extremist activities.

Van Valkenburgh (2021) focuses on a qualitative textual analysis of one such subculture, namely the subreddit "The Red Pill", which has the purpose to exchange *scientifically based* seduction strategies. Within this subreddit, supplementary readings in the form of 26 links are provided. Van Valkenburg directs his analysis towards these resources and thus analyses the underlying ideology. He finds that the subreddit has the purpose to provide a platform to exchange male *sexual strategies*. These are necessary as feminism has become the primary sexual strategy of women, putting them in the privileged situation to be able to choose the economically and biologically most ideal partner. Accordingly, the red pill subreddit aims to balance this wrong by giving "heterosexual men more power in pursuing individual sexual relationships within the existing system" (ibid., p. 89). Part of this strategy is men becoming more attractive to women "by mimicking alpha behavior and appearance" (ibid., p. 93).

Brooks et al. (2022) follow another approach and analyse whether incel activity on social media can be predicted by local socioecological circumstances like sex rations and income inequality. The reasoning behind that is that a male-biased sex ratio would suggest from a heterosexual point of view that some men remain unpartnered and thus might engage in incel activity. At the same time, high-income inequality could negatively affect chances of finding a partnership and thus increase incel activity. They tested their hypotheses on US commuting zones and identify levels of incel activity through a self-created dictionary based on incel jargon. Ultimately, they find that incels are more common in US commuting zones with male-based sex ratios and higher income inequality.

Scaptura and Boyle (2020) moreover analyse whether perceived stress by one's inability to fulfil gender roles or norms of masculinity is associated with a higher likelihood of fantasies about mass and gender-based violence. Similarly, they ask if what they call incel traits measured on a 20-item scale, capturing two dimensions, one revolving around exclusion and rejection and the other around hate and vengeance, are associated with violent fantasies about rape and using powerful weapons against enemies. They conduct an online self-report survey of 18- to 30-year-old heterosexual men in the US and found that men who are having trouble reaching up to expectations of masculinity or possess hostile incel traits, more frequently report fantasies about mass murder and rape.

Moreover, there is literature focusing on incel motivated violence. Hoffman et al. (2020) analyse whether the incel movement poses a terrorist threat by describing the community's ideological basis and the belief systems of its more extreme fringes. Furthermore, they provide a classification of incel violence and suggest four categories: clear incel-motivated terrorist attacks, attacks with mixed motives that evidence incel ideological influences, acts of targeted violence perpetrated by self-professed involuntary celibates, and ex-post-facto inceldom. Hoffman et al. (2020, p. 569) conclude that although "little-to-no coordination among the perpetrators" can be recognised "the homicidal intent that underpins this movement is undeniable, averaging almost eight fatalities per incident" (ibid.).

Similarly, Brace (2021, p. 4) debates whether incels are terrorists. He argues that within the incel community isolated individuals are advocating for an "incel uprising, a change in society's attitudes towards feminism, or policies that would result in women being forced to have sex with men". Accordingly, he claims that fostering an ideological goal and aiming at political change are constitutional aspects of violent terrorism. Consequentially, the attempt to start an incel rebellion can be characterised as terrorism (ibid.). At the same time, Brace also argues that the majority of incel-related violence is not driven by a higher ideological goal but out of personal revenge.

Apart from the general literature about incel violence, there is also literature focusing on single attackers, most prominently on the first incel terrorist Elliot Rodger who has been described from different angles as a mass murderer (White, 2017), a school shooter (Langman, 2016), a narcissistic personality (Allely & Faccini, 2017) and as a terrorist who conducted a clear incel-motivated terrorist attack (Hoffman et al., 2020). Moreover, Baele et al. (2021) analyse the world view behind the attackers of 2018, both Alex Minassian and Scott Beierle. The focus of the literature on Minassian is based on his published electronically recorded police interview immediately after his attack (A. Minassian, personal communication, 23 April 2018).

Rouda (2020) furthermore argues that the news media played a significant role in the escalation of incel violence over the last decade as they spread the ideology without critically highlighting or even truly questioning its seriousness. The media thus brought potential incels in touch with an explanation for their loneliness without emphasising the dangerousness. Additionally, the media sometimes misjudged the motivational basis and failed to recognise the assaults as what it was – an incel-motived act of violence. One prominent example is the German Hanau shooter of 2020 who was a publicly self-identified incel, but whose violence was primarily described as motivated by racism.

Additionally, there is a lot of research focusing on computational analyses of both the manosphere and the incel community. Farrell et al. (2019) analyse 6 million posts from 2011 to 2018 on Reddit and investigate the spread of information and misogynist ideas across seven online communities. Moreover, they created nine lexicons measuring different aspects of misogynism based on feminist critiques of language. These lexicons capture the following aspects: physical violence, sexual violence, hostility, patriarchy, stoicism (hardship due to lack of intimacy), racism, homophobia, belittling and flipped narratives. They find that hostility, stoicism and physical violence are the most popular categories across the seven communities. Regarding the evolution, they state that misogynist ideas, hostility and violence towards women online are firmly increasing over time across all communities within the manosphere.

Similarly, Horta Ribeiro et al. (2021) also trace the evolution of the manosphere from 2006 to 2018. They analyse 28.8 million posts from six forums and 51 subreddits. They use the above-mentioned dictionary by Farrell et al. (2019) to measure misogynism in a modified version with normalised counts. Additionally, they also used Google's Perspective API, which measures severe toxicity with the help of a convolutional neural network. Both tools were used to measure the evolution of toxic/misogynist content over time. They find that milder and older communities, such as Pick Up Artists (PUA) and Men's Rights Activists (MRA) are increasingly displaced by more radical, toxic and misogynist communities like incels or Men Going Their Own Way (MGTOW). Additionally, they find considerable user migration between communities.

Adding to this, Horta Ribeiro, Jhaver, et al. (2021) investigate how problematic online communities progress after they were forced to migrate to different platforms due to content moderation. They employ a discontinuity analysis and focus on two subreddits the r/The_Donald and r/Incel which have in common that they were banned and consequentially moved to stand-alone platforms. In both cases posting activity, active users and newcomers decreased after the migration. At the same time, they find that in the case of the subreddit the r/The_Donald, the remaining activity increased in toxicity, measured with the Perspective API, and radicalisation, determined on the basis of fixation and group identification. The latter was not significant for the incel subreddit. Overall, they conclude, that if platforms are banned it should be done rather earlier than later as migration is associated with halting community growth.

Focusing on incel-specific literature Jaki et al. (2019) analyse 65,000 posts from the incels.me forum in the period from 2017 until it was banned in 2018 and conduct a quantitative as well as a qualitative analysis of the rhetoric present in this discourse. Among other things they also ask how likely is it that the discourse in unmoderated incel forums leads to violent extremism. With the help of a small dictionary encompassing 10 offensive words, they find that about 30% of the posts are misogynist, about 15% are homophobic, and 3% are racist. Moreover, around 2% of the posts refer to actions of violence (kill, rape, shoot). Furthermore, they find that violence against women is accepted as desirable or even necessary and that incitements to violence, especially raping and killing women, are frequent. This concludes with the widespread extremist assumption that "the situation can only be improved by harming one of the out-groups (attractive men or women)" (Jaki et al., 2019, p. 20).

Likewise, Baele et al. (2021) also conduct a quantitative and qualitative analysis of the incels.me forum. They collected about 770,000 posts in the same period and focused on analysing the worldview shared by members of the incel community by especially focusing on elements associated with increased likelihood to engage in violence. Methodologically they focus on co-occurrence analysis as well as semi-supervised topic models. They address the in-group/out-group formation processes and argue that an extremist worldview is based on a "dichotomized, 'us versus them' thinking" leading to violence. Some groups are perceived as 'friends' (ingroups), others as 'enemies' (outgroups), with no possibility of a grey zone" (Baele et al., 2021, p. 1669). In the case of incels, the identified out-groups are *Alphas, Betas* and women as well as various derogatory descriptions of the latter including *Stacies, Roasties* (a judgmental term used for sexually active women) or *Bitches*. This incel-specific dichotomous process is moreover characterised by "a simplistic explanation of the positive ingroup suffering from the negative outgroups' nefarious actions, thereby pointing to violent solutions to restore the initial condition of the ingroup" (ibid., p. 1670).

Papadamou et al. (2021) analyse 6,500 YouTube videos shared on incel-related subreddits between 2005 and 2019 and focus on the evolution of the community using a dictionary with incel-related terms extracted from the Incel Wiki. Additionally, they investigate YouTube's recommendation algorithm. Overall, they find an increase in incel-related activity on YouTube encompassing both videos and comments over the last decade and conclude that YouTube is increasingly recognised as an incel platform to spread their misogynist ideology by its members. Additionally, when analysing the recommendation algorithm, they find that there is a 6.3% probability to end up with an incel-related video within five rounds of following the suggestions when starting with a random incel-unrelated video.

Moreover, there is literature focusing on automated detection of incel-related activity. For example, Hajarian & Khanbabaloo (2021) use roughly one million comments on Twitter and about 22,000 Facebook posts to identify incels by using an algorithm based on sentiment analysis. Their approach reaches an accuracy of 78.8%. They argue that incel detection on social media is crucial as it is the first step to stopping potential terrorists. Sang and Stanton (2020) follow another approach and base their training date for the automated detection on what they call incel hunter's critique, namely 18,000 screenshots collected from a subreddit specifically aiming at exposing incels. Their main finding is that they find a systematic relationship between the words used in the screenshot and the title attached meaning that titles as such can be used for efficient automated incel detection. Furthermore, they apply Plutchik's eight basic human emotions to the collected incel dataset and find that fear (17.6%) is the most frequent emotion, followed by anger (15.5%) and sadness (14.5%) while joy, trust, disgust, anticipation, and surprise were less frequent.

2.3 Theorising Twitter

This thesis is furthermore based on literature about Twitter. Murthy (2012) investigates Twitter in historical and broad sociological terms. He describes Twitter as a social media platform which he in turn defines as "a medium wherein 'ordinary' people in ordinary social networks (as opposed to professional journalists) can create user-generated 'news' (in a broadly defined sense)" (Murthy, 2012, p. 1061). Social is thus intended to explicitly exclude traditional media. Moreover, he argues that this "medium is designed to facilitate social interaction, the sharing of digital media, and collaboration" (ibid.). Additionally, he defines micro-blogging services, such as Twitter which he describes as

> an internet-based service in which (1) users have a public profile in which they broadcast short public messages or updates whether they are directed to specific user(s) or not, (2) messages become publicly aggregated together across users, and (3) users can decide whose messages they wish to receive, but not necessarily who can receive their messages; this is in distinction to most social networks where following each other is bi-directional (i.e. mutual) (Murthy, 2012, p. 1061).

Ultimately Murthy associates Twitter with a process of self-production and self-affirmation in the sense that "I Tweet, Therefore I Am" (ibid., p. 1062). Nevertheless, this also means that the self-confirmation via Twitter is continuously ongoing and needs constantly new input in the form of tweets. Apart from that, Twitter also offers the opportunity "to assert and construct the self which are contingent on a larger dialogic community" (ibid., p. 1063) and thus to position oneself within a broader community.

Similarly, Gruzd et al. (2011) debate whether Twitter-based relations can constitute a community, which is especially questionable in the case of Twitter as it is an asymmetric platform, meaning that the following relationship does not necessarily have to be mutual. They apply Anderson's (2016) theory of the Imagined Communities which are artificially constructed communities where members cannot possibly know everyone but share a sense of community to Twitter. Among other things an imagined community is characterised by the development of a common language. Gruzd et al. argue that this is true for Twitter through the usage of hashtags (#) to brand specific topics and facilitate exchange. They thus conclude that communities do exist on Twitter.

Additionally, Boyd et al. (2010) examine the practice of retweeting on Twitter. Before the retweet function was formally embedded in 2009 users needed to mark retweets with a specific syntax similar to this "RT @user 'message'". They provide a classification of ten motives to retweet which includes amplifying the tweets to a new audience, informing a specific audience, adding new content, highlighting one's presence as a listener, voicing agreement, validating other opinions, as an act of friendship, to refer to less visible content, to gain more attention, or to personally save them for later. Therefore, they argue that retweeting can be understood as both, "a means of participating in a diffuse conversation" (Boyd et al., 2010, p. 1) and as a mean "to validate and engage with others" (ibid.).

Another important topic related to Twitter is content moderation. Roberts (2019, p. 33) provides a definition of content moderation and describes it as an "organized practice of screening user-generated content posted to internet sites, social media, and other online outlets". Moreover, she differentiates between content moderation which takes place before or after the uploading procedure and between active and passive content moderation where the latter is only triggered in cases when content is flagged as problematic by other users. Additionally, her work focuses on the crucial working conditions for those employed as commercial content moderators, including constant exposure to the most disturbing content. According to her, those human content moderators remain essentially although algorithmic systems are increasingly employed to facilitate the handling of an ever-growing scope of potentially problematic content.

Gorwa et al. (2020) instead focused specifically on algorithmic and thus automated content moderation and investigate content moderation practices of Twitter, Facebook, and YouTube. They claim that Twitter relies on its Quality Filter which predicts whether tweets are of low-quality, spam, or automated. However, they also state that Twitter in general rather advocates for freedom of expression and thus prefers less drastic options like reducing the visibility of tweets over content removal.

Alizadeh et al. (2022) on the other hand investigate the salience of content moderation as a topic of discussion on Twitter. They argue that content moderation is "an increasingly contested practice, linked to fundamental political questions such as freedom of expression" (ibid., p. 1).

At the same time, they claim that there is "no established consensus exists regarding the nature of the problem, nor the appropriateness of potential solutions" (ibid., p. 3). They analyse over three million tweets from January 2020 to April 2021 and found that the salience of content moderation peaked in January 2021 after the storm on the US capital building in Washington DC and the subsequent ban of former US President Donald Trump.

2.4 Extremism and Radicalisation

Finally, this thesis considers the scholarly debate within the extremism and radicalisation field. After theorising extremism this section will focus on reviewing a definition of radicalisation, the conceptual difference between cognitive and behavioural radicalisation as well as the most prominent causes and models of the radicalisation process. Additionally, the lone-wolf terrorism concept will quickly be introduced.

Berger (2018, p. 9) sets out to find a definition of extremism. According to Berger extremism is a belief system that is characterised by two aspects. First, by the formulation of opposing in-group and out-groups. Accordingly, "in-groups and out-groups each represent an identity—a set of qualities that are understood to make a person or group distinct from other persons or groups" (ibid., p. 26-27). The second aspect of extremism is a "crisis-solution construct" (ibid., p. 37) which outlines actions alongside these identities. Consequentially, he arrives at the working definition where extremism "refers to the belief that an in-group's success or survival can never be separated from the need for hostile action against an out-group. The hostile action must be part of the in-group's definition of success" (ibid.). Consequentially, radicalisation refers to "the escalation of an in-group's extremist orientation in the form of increasingly negative views about an out-group or the endorsement of increasingly hostile or violent actions against an out-group" (ibid.).

When turning to radicalisation it is important to remark, that radicalisation studies generally heavily focus on jihadist radicalisation. That's why Agius et al. (2021, p. 1) argue that strategies to counter and prevent violent extremism have a "gender blind spot" and fail to acknowledge the threat arising out of misogynism and masculinism, especially in regards to the far-right.

Neumann (2013, p. 874) provides a classic definition of radicalisation and defines it as "the process whereby people become extremists". He, on the one hand, introduces the distinction between cognitive and behavioural radicalisation, which is essentially a distinction between the *endpoints* of radicalisation. Additionally, he also debates the relationship these two have regarding each other. Neumann himself defends the position that there is a causal link, and that battling cognitive radicalisation consequentially also prevents behavioural radicalisation.

Borum (2003) on the other hand doubts that there is a clear causal link between developing extremist ideas and taking extremist actions and thus advocates for a specific focus on behavioural radicalisation or what he calls "action pathways".

Furthermore, there is lots of literature about the causes of radicalisation. Although most of them focus specifically on jihadi radicalisation, they can provide some valuable insights into the general radicalisation process. Hafez and Mullins (2015) discuss the following four pieces of the "radicalization puzzle": personal and collective grievances, networks and interpersonal ties, political and religious ideologies, and finally enabling environments and support structures, most prominently the internet. Horgan (2008) identifies six risk factors that facilitate radicalisation. These are emotional vulnerabilities, dissatisfaction with the current state, identification with victims or personal victimisation, an understanding that violence is not inherently immoral, a sense of reward and finally kinships or social ties.

Moreover, there is some literature about models explaining how radicalisation occurs and proceeds. Moghaddam (2005) employs the metaphor of a staircase with six stages to describe the process of how an unsatisfied individual who experiences unfair treatment, identifies a target and source of misery and binds subsequentially with others that share the same grievances and eventually becomes a terrorist by conducting a violent act. Silber and Bhatt (2007) develop a model that explains jihadi radicalisation which involves the following four stages: pre-radicalisation, self-identification, indoctrination and jihadization. Borum's model (2003), which is also based on four stages, is in contrast formulated in a more general way. In the first stage "it's not right" an undesirable condition is identified. In the next stage, "it's not fair", the undesirable condition is framed as injustice. Furthermore, in the third stage, the "it's your fault" stage, a group or person is identified as responsible for the unjust situation and blame is attributed. By doing so a clear out-group is formed. In the last stage, those identified as responsible are labelled as evil ("you are evil") and this is accompanied by a dehumanisation which ultimately legitimises violence in the form of a terrorist attack against the out-group.
At the end of all these processes stands an individual that is ready to commit violence.

Clancy et al. (2021) present the terror contagion hypothesis which aims at explaining why members of an at-risk population take the path towards violent radicalisation and specifically focus on a social contagion process. They define terror contagion as "a form of social contagion spread through cultural scripts by incidents of mass violence" (ibid., p. 6). Their approach aims at explaining all sorts of violent radicalisation however they also explicitly argue that this was the case in the incel community after Elliot Rodger conducted the first incel-related terrorist attack. According to the authors, these cultural scripts not only include a template ideology, containing an explanation of grievances and an out-group responsibility as well -as violence as the identified method to right the wrongs, but also a template method that is suitable for contagion. In the case of incels, they identify the template method as mass violence through vehicular ramming and mass shooting. There are two important requirements. First, the at-risk population needs to be able to identify with the perpetrator's identity and secondly mass violence needs to lead to admiration rather than despite. If this is the case predatory mass violence can trigger subsequent mass violence, stimulated by media reporting about both template ideology and template method. Ultimately, "terror contagions can become self-perpetuating. Each subsequently completed act of mass violence furthers the replication and spread of cultural scripts sustaining the contagion of violent radicalization in the at-risk population" (ibid., p. 6).

A phenomenon which seems particularly important for incel radicalisation is the "lone-wolf-terrorism" which describes "terrorist actions carried out by lone individuals, as opposed to those carried out on the part of terrorist organizations or state bodies" (Hamm & Spaaij, 2017, p. 5). Moreover, "the lone wolf terrorist is typically someone who acts out of a strong ideological or religious conviction, carefully plans their actions, and may successfully hide their intentions from others" (ibid., p. 6). It is important to consider that "[l]one wolves do not operate in isolation, and their radicalization can be traced to various social networks" (ibid., p. 59) as is the case with incels who radicalise through an exchange of beliefs within the manosphere (Jaki et al., 2019).

2.5 Gaps in the Literature

In the final section of the second chapter, I am going to recap the most relevant contributions before highlighting some of the gaps in the literature. The scholarly debate on digital violence against women focuses on specific characteristics of "gendertrolling" (Mantilla, 2013) or "e-bile" (Jane, 2014) and its evolution towards encompassing more types of women. Additionally, the literature addresses how defending male dominance is the motivation for both offline and online harassment, namely "cybersexism" (Poland, 2016). Moreover, specific harassment campaigns (Hardaker & McGlashan, 2016; Massanari, 2017) or particular forms of online harassment (Powell & Henry, 2017) are discussed. Furthermore, real-life implications of online hate speech have been the focus of research (Filippo et al., 2015). However, it remains to be analysed whether toxic and misogynist content spreads further than less radical content.

Moreover, there is already a lot of especially qualitative literature exploring prominent themes within the incel-based ideology (Brooks et al., 2022; Daly & Reed, 2022; Ging, 2019;

O'Malley et al., 2022; Van Valkenburgh, 2021). Additionally, there is literature analysing the terrorist threat arising out of the inceldom with a specific focus on previous attacks as well as how individuals interact online and thus further radicalise (Brace, 2021; Hoffman et al., 2020; Langman, 2016; White, 2017). Besides, surveys of incels have investigated socio-economic demographics and predominant attitudes (Scaptura & Boyle, 2020; Speckhard et al., 2021). Furthermore, both the manosphere and the incel community online and their respective discourses are extensively analysed through computational analyses. However, the literature focused mainly on incel-specific forums (Baele et al., 2021; Horta Ribeiro, Blackburn, et al., 2021; Horta Ribeiro, Jhaver, et al., 2021; Jaki et al., 2019), Reddit (Farrell et al., 2019; Massanari, 2017) and YouTube (Papadamou et al., 2021) while leaving out Twitter.

Twitter in turn is analysed from a point of view where tweeting is a tool for self-reproduction (Murthy, 2012). Moreover, characteristics of Twitter communities (Gruzd et al., 2011) are discussed as well as the motivation behind retweeting (Boyd et al., 2010). Additionally, there is a lot of literature on content moderation in general (Roberts, 2019) and on Twitter (Alizadeh et al., 2022; Gorwa et al., 2020). However, to my best knowledge, this is the first study aimed at analysing how retweeting is used within the incel community.

Finally, this thesis builds on literature on extremism (Berger, 2018) and on radicalisation processes (Borum, 2003; Moghaddam, 2005; Neumann, 2013; Silber & Bhatt, 2007) as well as the causes of radicalisation (Hafez & Mullins, 2015; Horgan, 2008). Moreover, the contagion of terrorism (Clancy et al., 2021), as well as the phenomenon of lone wolf-terrorism (Hamm & Spaaij, 2017), are considered. However, all these points are primarily analysed from a jihadi point of view while there is far less literature on terrorist threats arising out of violent misogynism and the incel community (Agius et al., 2021).

Summarising the above-mentioned points, this thesis contributes to the scholarly debate by analysing whether toxic and misogynist tweets spread faster than less radical content within the Incel community on Twitter. The author, thus, adds to the scholarly debate by investigating incel activity on Twitter and contributes to shedding light on violent misogynism and incel radicalisation.

3. Theoretical Considerations

In this chapter, the author is going to introduce the theoretical considerations on which this thesis is based, namely the literature on information diffusion in an online environment and on the echo chamber effect.

3.1 Information Diffusion

The literature on information diffusion overall agrees that content with negative associations tends to spread further. For example, Jenders et al. (2013) applied the SentiStrength Algorithm to tweets, which returns two scores expressing the strength of positive and negative sentiments within this tweet. They find that tweets with a negative sentiment valence have a higher probability of being retweeted than tweets with both neutral or positive valence and argue that this is the case because negative experiences attract more attention and therefore increase the visibility of a tweet. Moreover, they find that tweets with stronger emotions, regardless of their valence, are more likely to be retweeted by using the divergence between the positive and the negative score. Similarly, Naveed et al. (2011) find that tweets with negative valence values measured with the Affective Norms of English Words dictionary are retweeted more often and thus conclude that bad news travel fast. Furthermore, Kim and Yoo (2012) applied the Linguistic Inquiry and Word Count (LIWC) dictionary and analysed the reply and retweet behaviour according to sentiments. They find that both types of Twitter interactions are positively correlated with the existence of negative sentiment words and negatively correlated with positive words. When looking at different categories of negative sentiments they find that angry tweets spark the strongest retweeting behaviour, followed by anxious tweets. Sad tweets, however, are negatively correlated with retweeting behaviour and thus spark fewer retweets. Likewise, Fan et al. (2016) analyse emotional contagion on weiboo, a Chinese platform similar to Twitter. They originally considered four categories: anger, joy, disgust, and sadness but quickly focused only on anger and joy as they found low probabilities of contagion for disgust and sadness. They concluded that anger is more contagious than joy meaning that angry messages spark more follow-up tweets or retweets than do joyous ones. They argue that this is the case because anger travels more easily along weaker ties than joy on social media. Additionally, Mathew et al. (2019) analyse the information diffusion dynamics of posts by hateful and non-hateful users on Gab, which is an interesting platform as Gab promotes "freedom of speech", renounces content moderation and thus portrays an ideal environment to study the spread of hateful content online. They find that content posted by hateful users tends to spread faster, and farther and reach a much wider audience as compared to the content generated by normal users. Moreover, they find that hateful users are more densely connected than non-hateful users.

Crockett (2017) provides an explanatory framework on why digital media exacerbate the expression of moral outrage. She argues that behaviour that condemns violations of moral norms is expressed more severely online as costs are reduced and the risk of retaliation is drastically lowered compared to real-life personal interactions. This is the case because the internet allows the formation of like-minded communities, or echo chambers, in which the danger of backlash is limited as the audience overall agrees on the topic. Additionally, communicating outrage while being hidden in an anonymous crowd further diminishes potential risks. Similarly, Makkar and Chakraborty (2020) find that hateful tweets are retweeted in a significantly higher magnitude compared to non-hateful ones. They argue that this is the case as hate speech is often characterised by the formation of echo chambers where "hateful contents are distributed among a well-connected set of users" (ibid., p. 4). The following sub-chapter is going to address the echo-chamber phenomenon in more detail.

3.2 Echo-Chamber Effect

The idea behind the echo chamber goes back to Sunstein (2002) and captures a group polarisation process where people selectively interact with like-minded others and thus form an ideologically aligned community, eliminating critical voices (Terren et al., 2021). As a consequence "people are hearing echoes of their own voices" (Sunstein, 2002, p. 177) which in turn reinforces the dominant opinion and ultimately leads to clustering and polarisation of segregated communities. The echo-chamber effect is driven by homophily, the human desire to interact with people expressing similar opinions where reaffirmation is associated with positive feelings, while diverging opinions cause emotional stress (Colleoni et al., 2014). It is argued that social media provide the ideal environment for echo chambers to flourish as they allow the formation of ideologically confirmed communities irrespective of geographic proximity. Additionally, the echo-chamber effect is believed to be exacerbated by platform-specific technological features like recommendation algorithms where previous activity influences and personalised the content display (Terren et al., 2021).

One prominent example is the personalisation algorithm of Facebook which Pariser (2014) describes as a "filter bubble" and a "you loop" that ultimately is able to change your identity through the options and views presented by the algorithm. He summarised this process as follows:

> Your identity shapes your media, and your media then shapes what you believe and what you care about. You click on a link, which signals an interest in something, which means you're more likely to see articles about that topic in the future, which in turn prime the topic for you. You become trapped in a you loop, and if your identity is misrepresented, strange patterns begin to emerge, like reverb from an amplifier (ibid., p. 70).

Similarly, Munn (2020) argues that some platforms are "angry by design" and "hate-inducing architectures" (p. 2). More concretely, he states that Facebook's news feed reinforces the expression of outrage based on the logic of enhancing engagement while YouTube's recommendation algorithm steers users toward more extreme videos. Both thus follow the assumption that toxic content drives engagement. He concludes that by promoting the spread of hateful content, these platforms contribute to its normalisation.

However, the empirical evidence for the echo-chamber effect remains contested (see Terren et al., 2021 for a comprehensive literature review). Therefore, Colleoni et al. (2014) contrast the two opposing views, namely the public sphere scenario with the echo-chamber scenario. According to the former, the internet and particularly social media enhance the exchange and confrontation of diverging opinions while allowing public dialogue. The latter in contrast suggests that the Internet reinforces already existing beliefs through a self-selection process. The focus of their analysis is thus whether Twitter is enhancing discussion among users with differing political opinions, or if it increases the exposure to like-minded people. They argue that Twitter is a particularly interesting platform to study political homophily as it can be a symmetric platform symbolised by mutual following as well as an information diffusion platform where following is asymmetric. Therefore, it can function as both "a social medium based on symmetric ties and as a newsy medium based on non-symmetric ties" (ibid., p. 320). Accordingly, they find that Twitter can enhance high levels of homophily and function as an echo chamber when focusing on its characteristic as a social medium. But at the same time, Twitter can also lead to lower levels of homophily and thus foster public exchange when focusing on its characteristic as a news medium. Therefore, they partially confirm the echo-chamber effect on Twitter when it is used as a social network and not as a news platform.

When echo-chambers form, they can have crucial consequences by fostering social polarisation and the creation of estranged communities. This is also true for the manosphere and the incel community. Accordingly, Hoffman et al. (2020, p. 575) claim that "[b]y establishing ideologically cohesive echo chambers, social media unites disparate individuals separated by background and geography and offers a networked universe and common purpose". Moreover, the echo-chamber effect inhibits a dangerous radicalisation potential. According to Jaki et al. (2019, p. 1) "'echo chambers' […] where like-minded people share disparaging views, can be a catalyst for radicalization". Similarly, Baele et al. (2021, p. 1686) state that:

> The role of the Internet in enabling the formation and radicalization of this community through echo-chamber dynamics is evident: without a way to relate and discuss, these individuals would have had no way to recognize themselves as "Incels" and learn the culture and particular idiom that cements the Incel worldview.

To summarise it can thus be assumed that the echo-chamber effect inherits a dangerous radicalisation potential within the incel community. The following subchapter is going to introduce the hypotheses on which this thesis is based by investigating what influences tweet diffusion in the incel community on Twitter.

3.3 Formulation of the Hypotheses

When considering the above-mentioned arguments, it can be seen that there is an overall scholarly consensus that negative, especially angry, or hateful content diffuses further online. This is the case as echo-chambers form and reinforce pre-existing beliefs while radicalisation is enhanced. Applying this theoretical knowledge to the incel community on Twitter the author hypothesises that similar behaviour can be observed. As established by the literature, the incel community can be characterised as toxic and enabling radicalisation (see for example Horta Ribeiro et al., 2021; Baele et al., 2021). The author thus expects a positive relationship between the toxicity and the number of retweets of a tweet and thus hypothesises that:

H1: T*he more toxic a Tweet, the more often the Tweet is retweeted within the incel community on Twitter.*

The toxicity of tweets will be determined with the Perspective API measuring severe toxicity (*Perspective API*, 2018)

Furthermore, a second dimension is added. Apart from being a toxic community, the incel community is particularly characterised by its misogynism (see for example Farrell et al., 2019; Jaki et al., 2019). Therefore, a second hypothesis is added to investigate the relationship between misogynism and retweeting behaviour. This is done for two reasons. First, to verify whether toxic tweets are also considered misogynist and the other way around and thus to verify the measurement. On the other side, as mentioned above, it can be expected that within the incel community, which is particularly characterised not by its toxicity but by its misogynism, the latter can be expected to particularly drive retweeting behaviour. The author thus expects a positive relationship between the percentage of misogynist words in a tweet and the number of retweets and thus hypothesises that driven by an echo-chamber effect:

H2: *The higher the percentage of misogynist words in a Tweet, the more often the Tweet is retweeted within the incel community on Twitter.*

The percentage of misogynist words will be determined with a dictionary measuring misogynism that was developed by Farrell et al. (2019), which as well will be discussed in more detail in the next section discussing methodological considerations.

4. Methodology

In this chapter, I am going to address methodological considerations. After introducing the data, I will describe and define the variables and discuss the employed method as well as potential limitations.

4.1 Data

The data used for this analysis was gathered with the help of the R package rtweet (Kearney et al., 2020) over almost seven weeks from March 21, 2022 to May 06, 2022 with the Twitter API (*Twitter API*, 2020). The time frame was chosen to represent any ordinary period. Twitter provides two options to collect public data either through its streaming (which allows real-time live data collection) or REST API. The author decided that the REST API, which returns data from the past six to nine days, based on defined search queries, is sufficient for the purpose of this thesis.

Collecting and analysing data from Twitter brings certain ethical and legal issues, most prominently consent and privacy (see for example Gold & Department of Computer Science at UCL, 2020). Consent from Twitter to use the data within agreed-on terms and conditions is provided automatically when the application for the Twitter developer account is granted as this application also includes detailed descriptions of the intended purposes. Moreover, consent from users can be assumed indirectly as Twitter explicitly states during the registration process that the data can be used for research purposes. Additionally, users, when tweeting, intentionally decide to publicise their content. Privacy, however, is protected to the most possible extent within my analysis as tweets are generally only displayed in an aggregated form with only two exceptions. In these cases, the username is removed and only the content of the tweet is kept, reaching almost full anonymity. Finally, Twitter allows the publication of tweets and user IDs for replication purposes.

The data collection process was repeated every morning during the collection period and subsequently, duplicates were detected and deleted and only those duplicates with the highest retweet count and thus the assumed most recent ones were kept. The collection was based on 12 keywords which are prominent within the incel community. The keywords were pre-selected with the help of two glossaries, the Incel-Wiki[4] and Tim Squirrel's Glossary[5], and by scrolling

[4] Wiki Incel Glossary. Retrieved from: https://incels.wiki/w/Incel_Glossary, checked 09.06.2022.

[5] Squirrel. A definitive guide to Incels part two: the A-Z incel dictionary. Retrieved from: https://www.timsquirrell.com/blog/2018/5/30/a-definitive-guide-to-incels-part-two-the-blackpill-and-vocabulary, checked 09.06.2022.

through Twitter. Subsequently, the author researched on Twitter whether the keywords seemed to be used within the incel context. As a result, many potential keywords were excluded. This was for example the case when keywords were used in a rather general context and not incel-specific enough (e.g.: foreveralone, itsallover) or used to talk about incels but not used within the community as was the case with the keyword incel. In Table 1 the selected keywords are listed and briefly explained.

Keyword	Explanation	Keyword	Explanation
awalt	abbreviation for "all women are like that" and used to generalise female behaviour	**femoid**	derogatory term for women composed by female and -oid (e.g.: android)
betamale	average male as opposed to and subordinated by alpha males	**goingER**	following Rodger Elliot's example and committing mass murder for the incel cause
betafags	derogatory term for beta males	**incelrevolution**	revolution by the incel community to express revenge
betauprising	used to describe an incel revolution with the aim of revenge	**mgtow**	abbreviation for "Men Going their Own Way" describing a belief to be better off without any interaction with women
blackpill	accepting the fatalistic truth that society is fundamentally against men and in favour of women and nothing can be done about it	**redpill**	accepting the truth that society is fundamentally against men and in favour of women
bluepill	choosing to believe a comforting lie about women and society	**trueincel**	true incel as opposed to a fakecel, someone who is pretending to be an incel and could be successful with women

Table 1: Keywords for Data Collection

Based on these 12 keywords a total of 52,927 tweets were collected. However, the author decided to include only English tweets for the analysis and thus excluded all the non-English tweets based on a variable in which Twitter automatically identifies the language of the tweet. Moreover, tweets by the seven most active users, with 296 to 1,245 tweets each, were manually removed to prevent biases. Particularly because these users were included as one of the keywords was a part of their username and reviewing a subset of their respective tweets revealed that the majority of the tweets were not related to incel-specific topics. Ultimately, after having a look at the text data, the author also decided to exclude tweets with less than 20 characters based on variable provided by Twitter that indicates the length of the tweets. This was done as these short tweets seemed inconclusive and lacked content.

The Twitter data does not only include a unique ID for the user and the tweet itself as well as the text but a total of 90 variables. However, most of these variables are not considered. Eventually, a total of 19,359[6] tweets by 10,789 users were kept for the analysis. The users contributed on average roughly two posts (1.79), while 8,681 users were only included once and 1,159 twice. Moreover, only 12 users contributed more than 100 tweets, with the maximum being 273. The author believes that the combination of these should shield against grave biases.

Due to the design of the data collection process, a within-community analysis follows. It was a specific goal to capture a snapshot of the incel discourse on Twitter. Consequentially, only variation within an already rather radical context can be investigated. The goal is thus not to draw any comparison to other communities on Twitter but to analyse variations of radicalisation, be it toxic or misogynist, within an already radical environment of the incel community on Twitter.

Figure 1 displays the tweet distribution over time. It can be seen that the time span of collected tweets extends the collection period. This is the case as data from up to 9 days prior to the first collection round can be gathered. On average 345.7 tweets were added every day, with considerable lows on the first (9) and last day of the data collection process (137). The most tweets (586) were added on 29.04.2022 followed by 518 on 19.04.2022.

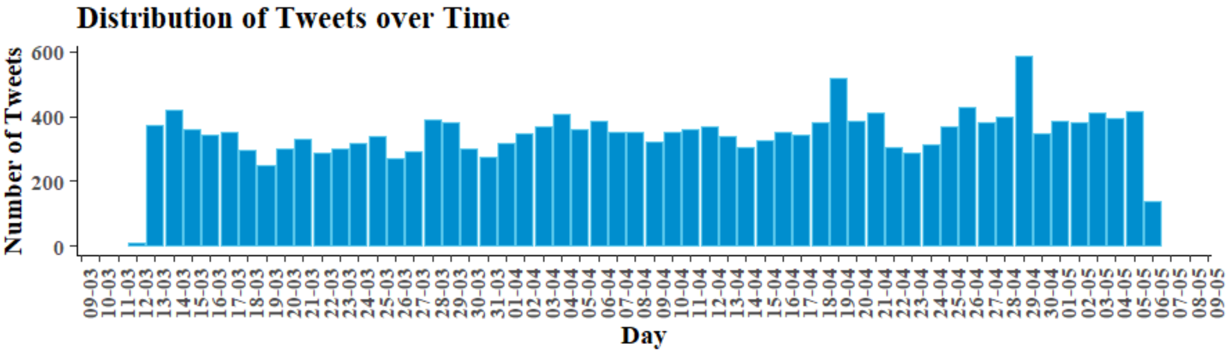


Figure 1: Distribution of Tweets over Time

When having a deeper look into the content of the tweets, it is not surprising, that the search keywords are among the most frequent words, immediately followed by the word women. Table 2 displays the frequencies of the ten most common words while Figure 2 illustrates words which are included in the dataset more than 200 times in a word cloud with those being more frequent, displayed in bigger fonts.

[6] Tweet IDs for replication can be found here: https://drive.google.com/drive/folders/1xKz7pQLv0gljqohgz5w2_xZDllxF8tUK?usp=sharing.

Rang	Term	Frequency	Rang	Term	Frequency
1	redpill	6997	6	just	1773
2	mgtow	3993	7	like	1766
3	blackpill	3742	8	can	1186
4	men	2115	9	get	1158
5	women	1974	10	people	1153

Table 2: Most frequent Words

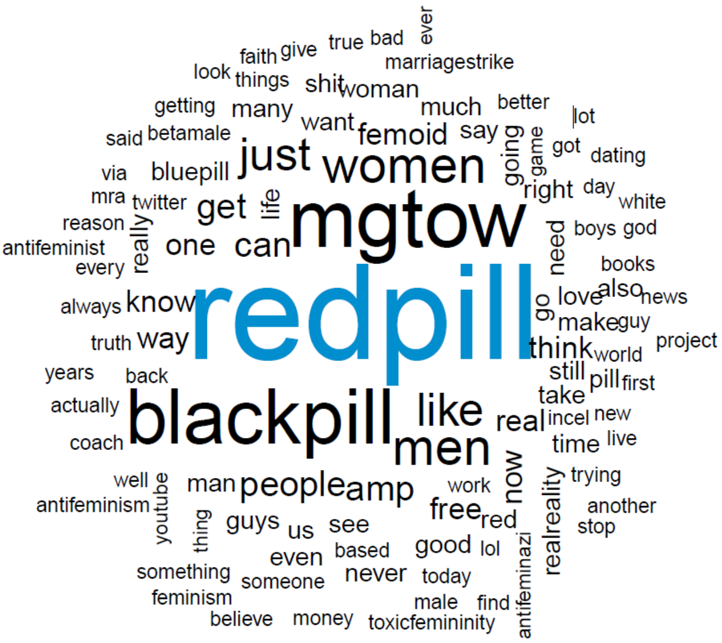


Figure 2: Word Cloud with most frequent Words

4.2 Variables

Now the dependent, independent and control variables on which this analysis is based are introduced and defined. The dependent variable measures the information diffusion on Twitter based on the number of retweets. As argued by Boyd et al. (2010) retweeting behaviour can have various motivations including, for example, to diffuse tweets to a new audience, to voice agreement or to validate other opinions. The collected Twitter data includes a variable *retweet_count* which provides an integer "number of times this tweet has been retweeted".[7] As mentioned above duplicates tweets were removed so that the assumed most recent value with the highest retweet count was kept. Figure 3 provides an overview of the distribution of retweets.

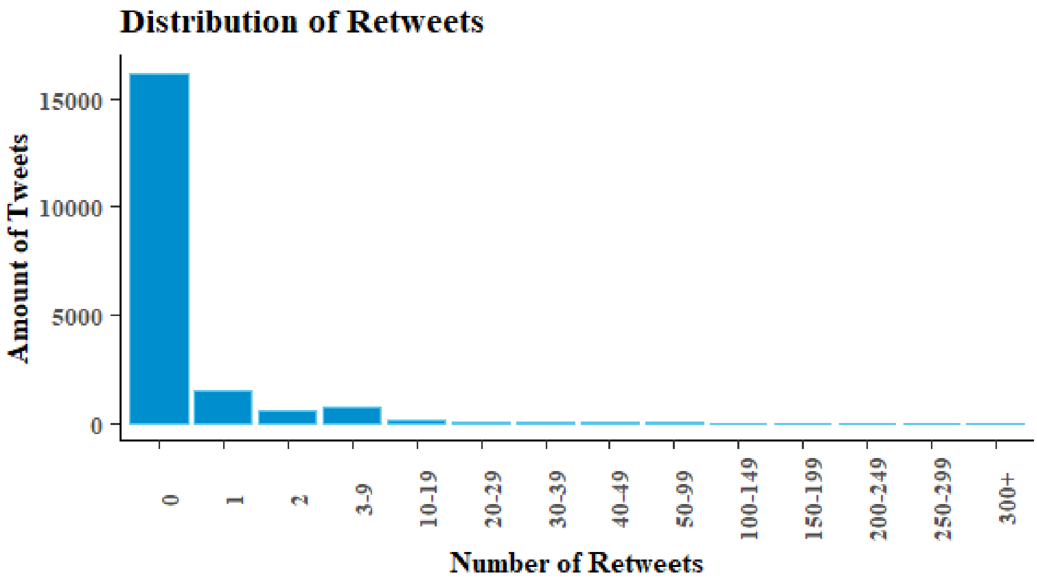


Figure 3: Distribution of Retweets

Of all the collected tweets 16,190 were not retweeted at all. 1,500 tweets have been retweeted once, 541 have been retweeted twice and 765 three to nine times. The maximum a tweet has been retweeted is 1,423 times, followed by 609 and 581 retweets. Ultimately, in order to approach a normal distribution and to limit the influence of outliers, logarithmic retweet count numbers were used. This led to the distribution displayed in Figure 4.

The first independent variable measures the toxicity of a tweet with the help of the Perspective API (*Perspective API*, 2018). The Perspective API is a product offered by Google and Jigsaw which is based on machine learning and aims at supporting content moderation by flagging toxic or hateful content in comment sections or forums and thus contributes to creating a safer environment. Prominent users of the Perspective API are for example the New York Times or Reddit (ibid.). Perspective's most prominent attribute is *toxicity* which is defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" and is available in 17 languages.[8]

[7] Twitter Developer Platform. Twitter Variables. Retrieved from: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet, checked 09.06.2022.
[8] Perspective API. Attributes and Languages. Retrieved from: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages, checked 09.06.2022.

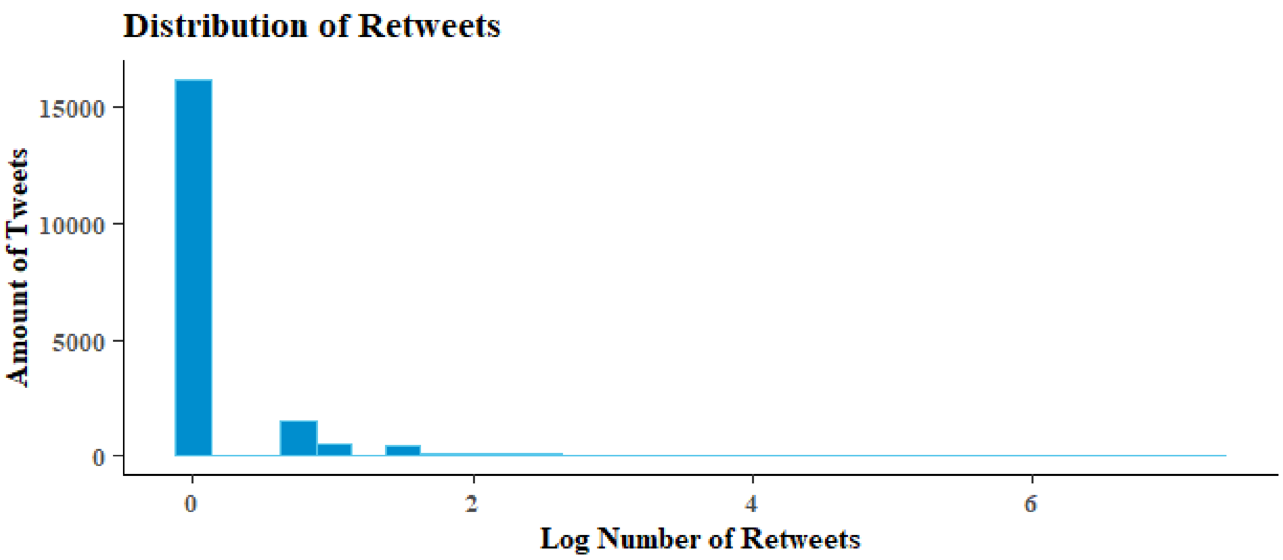


Figure 4: Log Distribution of Retweets

The author, however, decided to use the *severe toxicity* attribute which is defined as “a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective” as this attribute is less sensitive to weaker forms of toxicity like positively using curse words.[9] The Perspective API is not only used commercially but has also been employed in research. For example, Adewale Obadimu et al. (2019) used the toxicity attribute of the Perspective API to analyse comments posted on pro- and anti-NATO channels on YouTube. Moreover, Horta Ribeiro et al. (2021) applied the severe toxicity attribute to analyse the evolution of toxicity in the manosphere. Additionally, Horta Ribeiro, Jhaver, et al. (2021) used the Perspective API to trace levels of toxicity before and after platform migration due to content moderation. The performance of the Perspective API has been assessed by previous work. Rajadesingan et al. (2020) verified with a small sample of 100 comments, that in political subreddits, the toxicity attribute of the Perspective API achieves similar results as a collection of eleven human labellers. At the same time, Zannettou et al. (2020) found that also on a subset of 100 comments the severe toxicity attribute of the Perspective API outperforms the Hate Sonar classifier. Additionally, Hosseini et al. (2017) analysed the weakness of the Perspective API by demonstrating that it is vulnerable to manipulation, especially through misspelling or adding punctuation between the letters of abusive words which lead to significantly lower toxicity scores. However, taking the volume of analysed tweets and assuming that the incel community on Twitter has no intention to hide the real extent of toxic language, the author decided to neglect this risk and proceed with the

[9] Perspective API. Attributes and Languages. Retrieved from: https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages, checked 09.06.2022.

Perspective API. The first independent variable thus portrays a toxicity score for every tweet indicating the likelihood of this tweet being severely toxic, ranging from 0 to 1 with higher values indicating higher toxicity. This score is determined with the R package peRspective (Votta, 2021). With the help of the Perspective API 18,826 were returned successfully while the remaining 533 received an error message due to unsupported and falsely recognised languages. The mean toxicity score is 0.052 thus indicating a low overall toxicity within the collected tweets. Moreover, Table 3 displays descriptive statistics of the toxicity score.

Descriptive Statistics Toxicity Score

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Toxicity Score	18,826	0.052	0.102	0.00001	0.003	0.045	0.930

Table 3: Descriptive Statistics of Toxicity Score

Figure 5 displays the distribution of the toxicity score while for illustration purposes, Table 4 presents a few randomly selected tweets and their respective toxicity scores.

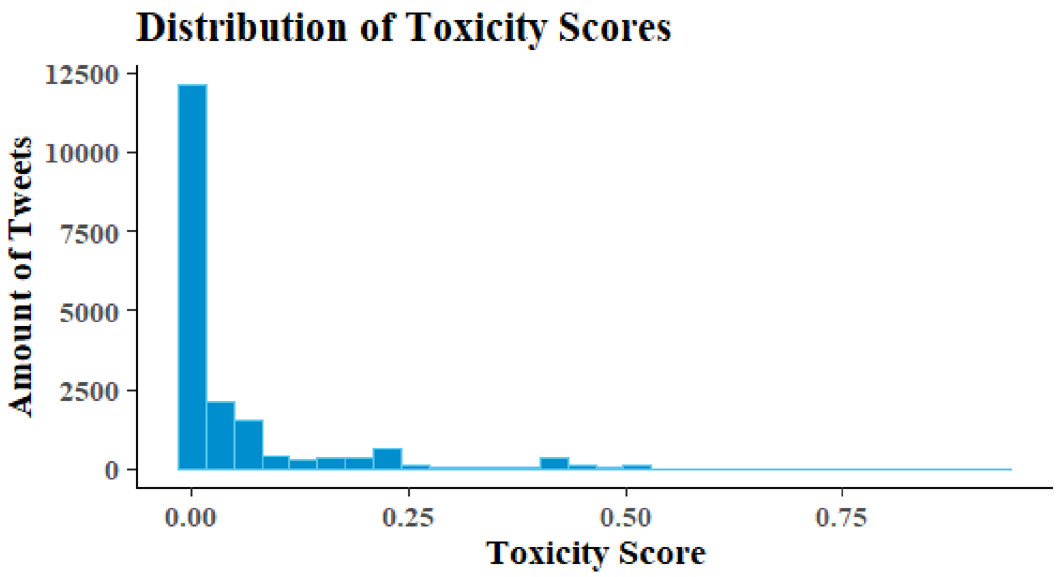


Figure 5: Distribution of Toxicity Scores

Tweet	Toxicity Score
This reads like incel blackpill ideology with the words switched	0.250
blackpill dudes deserve to be bullied	0.518
ALL WOMEN HAVE SHIT TASTE #INCEL #BLACKPILL #ITSOVER	0.694
@RedPill_Belgium Fucking muslims. Do a Putin on them	0.749

Table 4: Selection of Tweets and Toxicity Scores

When looking at the average toxicity scores per day (Figure 6), no apparent increasing or decreasing toxicity trend can be identified within the short study period. However, there is some variation between 0.03 and 0.06 and low average daily toxicity scores seem to appear on weekend days.

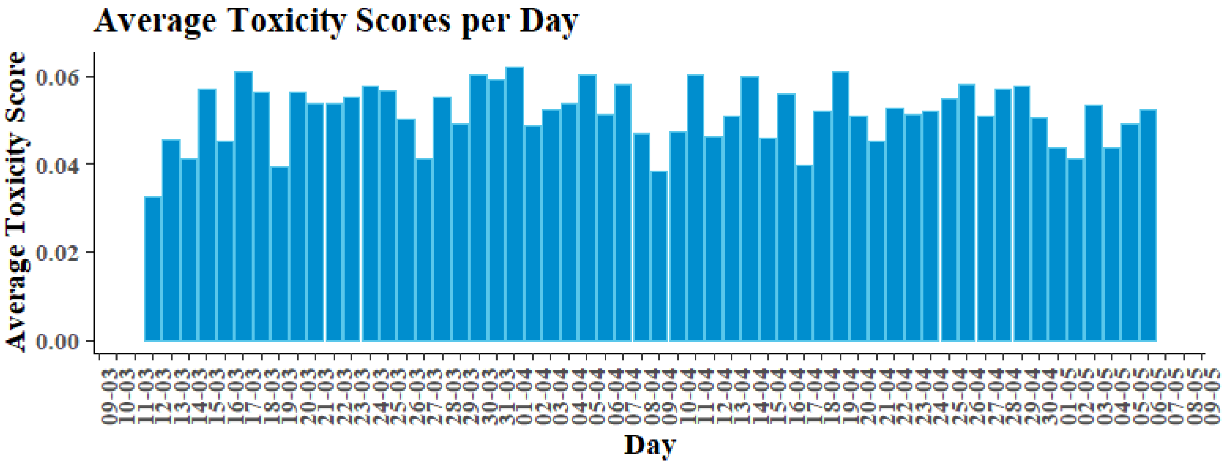


Figure 6: Average Toxicity Scores per Day

Apart from this continuous toxicity variable ranging from 0 to 1, a binary toxicity variable was created. This decision was made as a reaction to the non-normal data distribution of the toxicity score as the skewness of the distribution can crucially influence and bias the results. However, as no tweet received a toxicity score of truly 0 a cut-off value had to be chosen. I decided to use the first quartile value (0.003) and accordingly coded everything smaller than 0.003 as not being toxic (0) and everything greater or equal to 0.003 as toxic (1) which led to a distribution of 4,663 tweets being coded as non-toxic and 14,163 as toxic.

The second independent variable measures the misogynism of a tweet with the help of a dictionary developed by Farrell et al. (2019).[10] Misogyny is derived from the Greek word *misoginìa* which is a composition of *miso-* (to hate) and *-gyne* (woman) (Frenda et al., 2019).

[10] Github. Misogynism dictionary. Retrieved from: https://github.com/miriamfs/WebSci2019/blob/master/Lexicon.txt, checked 09.06.2022.

Farrell et al. (2019, p. 3) themselves define misogyny “not only as behaviour that objectifies, reduces, or degrades women but also as the exclusion of women, manifesting itself in discrimination, physical and sexual violence, as well as hostile attitudes toward women”. Moreover, Anzovino et al. (2018) describe five components of misogynism: discrediting women, spreading stereotypes and objectification, intimidating women through sexual harassment and threats of violence, demonstrating male dominance and ultimately derailing which legitimises abuse. Since I use the dictionary of Farrell et al. (2019) it is only logical to follow their understanding of misogynism. The dictionary captures the following nine dimensions of misogynism: belittling, flipping the narrative, homophobia, hostility, patriarchy, physical violence, racism, sexual violence, and stoicism. However, I decided to exclude the categories racism (around 700 words) and homophobia (around 300 words) as they seemed too broad for the research purpose of this thesis and measured other kinds of hate speech. Moreover, racism was by far the biggest category while homophobia was the third biggest category, thus threatening to essentially measure the wrong phenomenon. Additionally, I excluded the keywords on which the data collection was based and bigrams as the analysis only focuses on single words. Eventually 440 terms remained in the dictionary.

Before applying the dictionary to the tweets, the column containing the text of the tweets was extracted, turned into a corpus, uppercase letters were turned into lowercase letters and English stop words as well as punctuation characters (including URLs, symbols, separators and numbers) were removed. Subsequently, the dictionary was applied to the data and the number of hits was normalised by the overall number of words in the tweet. This provided us with a misogynism score ranging from 0 to 1 and measures the percentage of misogynist words in a tweet. The second independent variable thus portrays a misogynism score for every tweet indicating the percentage of misogynist words in the tweet. The mean misogynism score is approximately 1% and 16,917 tweets do not contain any of the dictionary words. Table 5 provides further descriptive statistics and Table 6 displays some demonstrative tweets and their respective misogynism scores with the dictionary matches written in italic letters.

Descriptive Statistics Misogynism Score

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Misogynism Score	19,359	0.010	0.035	0	0	0	0.50

Table 5: Descriptive Statistics Misogynism Score

Tweet	Misogynism Score
hoes really *hate* niggas [sic!] w/ redpill mindsets ..be catching the lies soon as they open they mouth	0.166
I hate the *blackpill* and I *hate* being short and I *hate* how society treats short men. I just want a normal height so I can live a normal life.	0.176
blackpill is for fucking losers. It's *cope* because you're too lazy work on what you can so you just make excuses and cry like a little *bitch* about *hypergamy*. Maybe if you hit the gym instead of making excuses for why you have no *bitches* you wouldn't be bitchless	0.208
Lifting can't fix *beta*. Game can't fix *beta. Beta* is beta. #BlackPill	0.400

Table 6: Selection of Tweets and Misogynism Scores

This short demonstration of randomly selected tweets already indicates a potential weakness of this approach. While the last tweet is supposed to be the most misogynistic tweet it is in fact not at all. At the same time the second and the third tweet indicate that expressions of hate do not always seem to be directed at women but also at incels themselves.

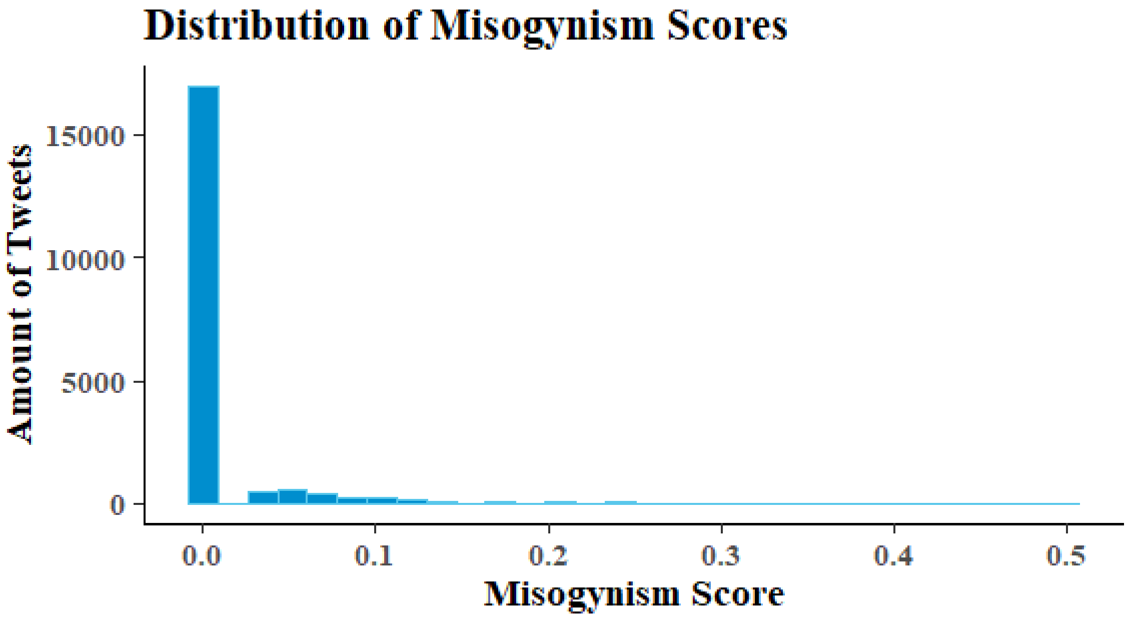


Figure 7: Distribution of Misogynism Scores

Figure 7 shows the distribution of the misogynism scores and Figure 8 the average misogynist scores per day during the study period. The highest average of 2.7% was reached on the first day of the data collection. However, on this day only nine tweets were collected, and the average score is thus neglectable. Like for the toxicity score, no overall trend can be seen. The averages vary between 0.05 and 0.14 when neglecting the outlier on the first day. In contrast to the toxicity score low averages do not seem to appear on specific days.

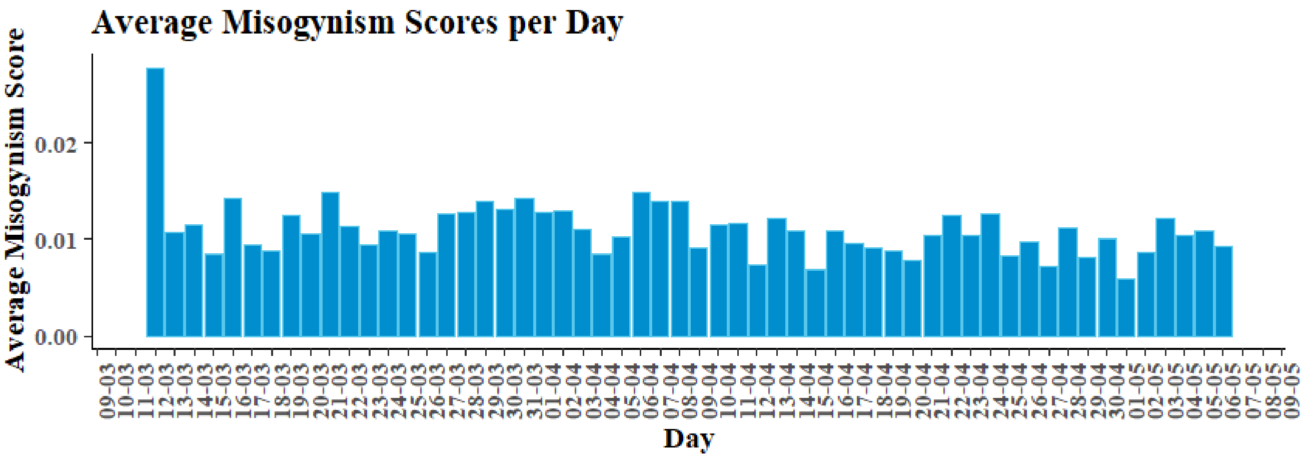


Figure 8: Average Misogynism Scores per Day

Additionally, Figure 9 provides an overview of the most frequent dictionary matches. While the dictionary is obviously in its nature biased towards misogynist and thus more negative and hateful words, it remains astonishing that strongly vicious words like hate, hit, destroy, hurt, kill, rape and beat are among the 20 most frequent words. Similarly, to the toxicity score a binary variable measuring misogynism was introduced, as the misogynism score as well is not normally distributed. Due to the construction of the misogynism score, tweets can have the value 0, which is the case if there are no dictionary matches at all. Therefore, the variable was coded in such a way that a misogynism score of 0 also translated into being non-misogynist (0) while tweets containing a higher value than 0 were coded as misogynists (1) which resulted in 16,917 tweets being coded as non-misogynist and 2,442 coded as misogynists. When analysing the correlation between the toxicity score and the misogynism score (Figure 10) only a weak correlation (r= 0.211) can bet detected.

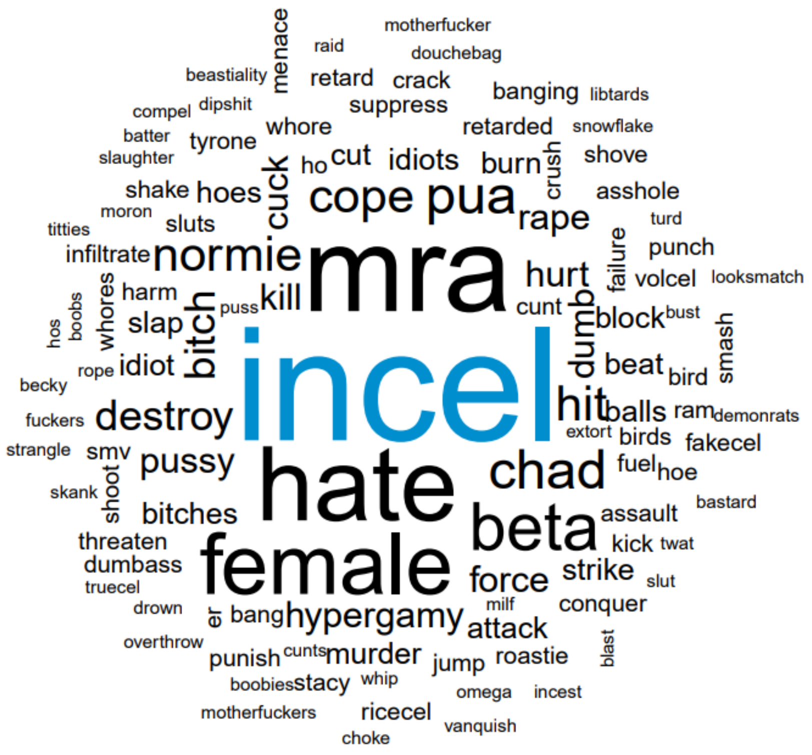


Figure 9: Word Cloud with most frequent Dictionary Matches

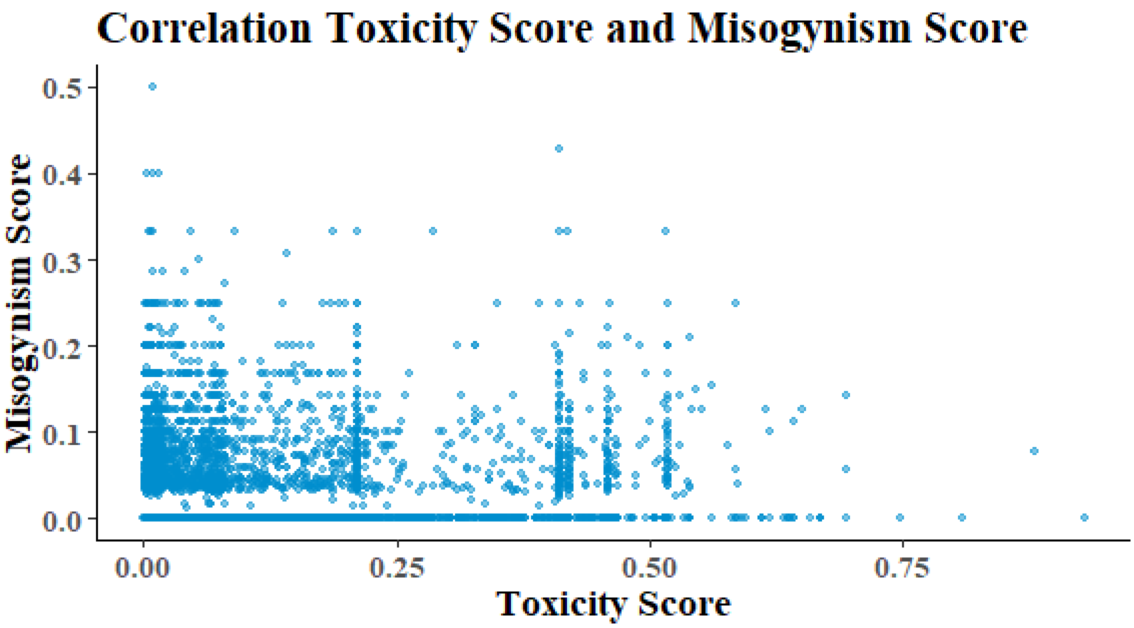


Figure 10: Correlation Toxicity Score and Misogynism Score

This indicates that the two measurements of the toxicity with the help of the API and the misogynism determined with the dictionary approach, do have a conceptually different focus, and thus capture different aspects.

Apart from the dependent and the two independent variables, the author decided to include three control variables. After reviewing the literature on tweet diffusion, the following three non-context related factors were identified as potential confounders, which can influence both the dependent and the two independent variables and thus have to be conditioned for: the presence of hashtags (Comarela et al., 2012; Jenders et al., 2013; Naveed et al., 2011; Suh et al., 2010), the presence of URLs (Comarela et al., 2012; Jenders et al., 2013; Naveed et al., 2011; Suh et al., 2010) and the number of followers (Jenders et al., 2013; Suh et al., 2010). These three factors were all positively associated with higher retweetability. Moreover, they could potentially influence the toxicity as well as the misogynism of a tweet. The presence of URLs is theorised to be related to the spread of presumed interesting articles, web pages or videos (Suh et al., 2010). Retweeting of tweets containing URLs can thus intuitively be understood as acknowledging the contained information as worthy of spreading. Hashtags on the other hand facilitate the categorisation of tweets into pre-defined topics and thus help to spread topic-related information (Firdaus et al., 2018). Hashtags, therefore, facilitate retweeting by making tweets belonging to a particular topic easily available. Both hashtags and URLs can influence the toxicity or misogynism score if they include terms which are understood to be either misogynist or toxic. Furthermore, having more followers and thus a larger audience to which the tweets are exposed, makes it more likely for tweets to be retweeted (Suh et al., 2010). A higher number of followers in turn might lead to higher toxicity and misogynism scores as bigger accounts might feel pressured to deliver more radical content in order to keep expanding the audience.

The necessary information for the control variables is included in the dataset. The variable *hashtags* either contains a vector with the employed hashtags in this specific tweet or display NA if no hashtags have been used. Based on this information, the author coded a dummy variable where 1 corresponds to the presence of hashtags in the tweet and 0 corresponds to no hashtags present in the tweet. Similarly, the variable *urls_url* either displays the employed URLs or NA if no URLs were used. Accordingly, the author coded a dummy variable where 1 corresponds to the presence of URLs in the tweet and 0 to no presence of URLs. Finally, the number of followers could be directly extracted from the variable *followers_count*. For the number of followers, logarithmic numbers were used in order to limit the influence of outliers.

4.3 Method

To test the above-presented hypotheses and to answer the research question, the author decided to rely on Ordinary Least Square (OLS) regression as a methodological tool. For both independent variables, a total of four models were constructed. First, a linear model only including the independent variable, the toxicity and misogynism scores respectively. Then a linear model using the binary versions of the independent variables is calculated. Subsequently, in the models three and four a multiple regression was employed in which the influence of URLs, hashtags and the number of followers of the account are controlled for. One important assumption of multiple regressions is that there is no perfect multicollinearity between any of the independent variables (see for example Field et al., 2012). To verify that this is the case, I created correlation matrixes for both hypotheses (Figure 11and Figure 12). None of the variables has a very strong correlation which is higher than 0.8. This however has one exception, the correlation between both versions of the misogynism score, which is 0.81. But the usage of these variables' alternates, and they are never included at the same time.

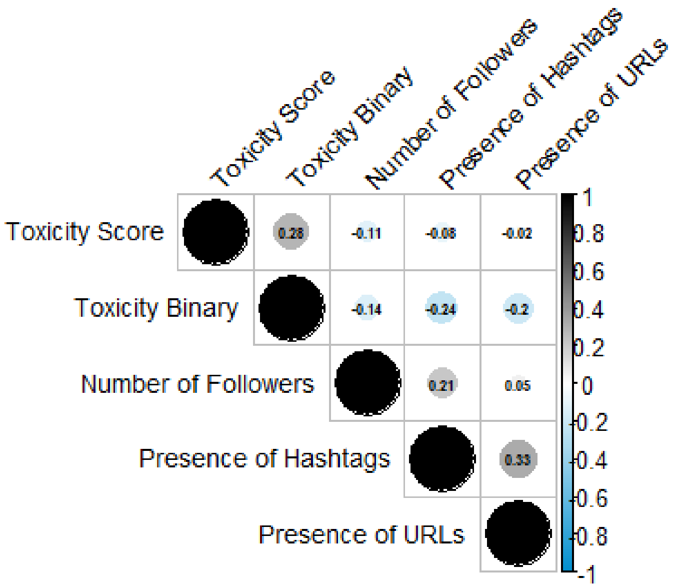


Figure 11: Correlation Triangle Independent Variables H1

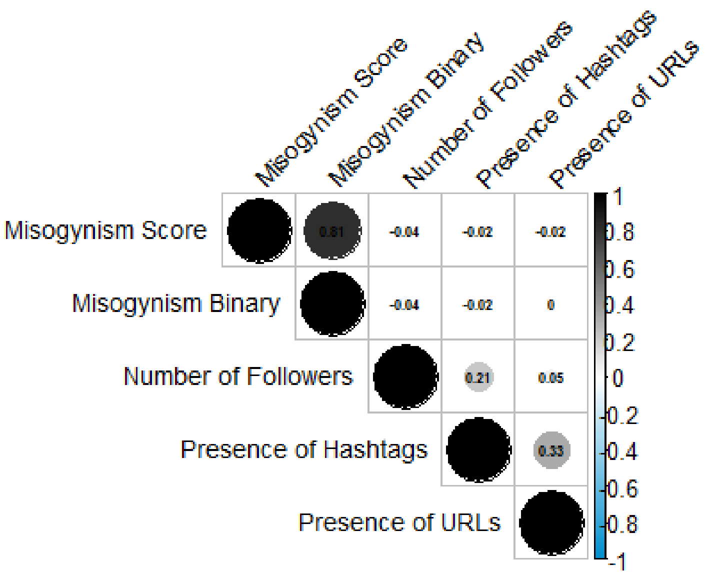


Figure 12: Correlation Triangle Independent Variables H2

Additionally, as the last control variable refers to user-specific characteristics and not the tweets as such, the tweets cannot be described as independent from these variables. Therefore, a key assumption is violated (Gelman & Hill, 2007). Consequentially, a fixed-effects model with user-fixed effects is employed in the last two models. The Hausman test (see for example Greene, 2012) confirms that the fixed effects model is the preferable and most consistent option ($p < 0.05$). Fixed effects models can be used with any kind of multi-level data – as in our case tweets nested in users and are characterised by including group-specific, i.e., user-constants into the model. More concretely, unobserved heterogeneity can be controlled by subtracting means across users (Brüderl & Ludwig, 2014).

4.4 Potential Limitations

Before reporting the results of the analysis a few potential limitations of the research design and the method have to be discussed. Firstly, it is important to mention that computational analyses based on textual data can inherit certain pitfalls. Grimmer and Stewart (2013) argue that although automated methods are necessary when considering the sheer amount of produced language, these methods necessarily produce "incorrect models of language" (Grimmer & Stewart, 2013, p. 268). They further elaborate that "any one sentence has complicated dependency structure, its meaning could change drastically with the inclusion of new words, and the sentence context could drastically change its meaning" (ibid., 270). Essentially, this should remind us to be careful and to validate the results of the automated methods as thorough as possible.

Additionally, the self-collected data set based on prominent keywords inherits the risk of not essentially capturing the inceldom. This risk is exacerbated by Twitter, which compared to for example Reddit does not have pre-defined communities (see for example Farrell et al. 2019, p. 1). Instead, communities form around the usage of specific hashtags or keywords. However, those keywords risk being used equally by members of the community as well as users talking about the community. Moreover, the keyword search did not only include the text of the tweets but also the username. Consequentially, tweets were sometimes included because of the username rather than on the basis of the content of the tweet. Therefore, also non-incel-related tweets were included in the data set. Moreover, the data set might be biased as the most radical content could have been deleted and thus not been captured by the data collection process.

Furthermore, Farrell et al. (2019) discuss two limitations of using dictionaries. First dictionaries tend to leave out certain important terms and thus lack completeness. Moreover, dictionaries cannot capture contextual details and thus risk error-proneness. One issue arising out of missing context factors that could already be seen in the small demonstration of tweets above is that both the toxicity score and misogynism score are not able to detect the object or direction of hate. Consequentially, sometimes tweets receive high toxicity or misogynism score merely because hateful words are included although the hateful content of the tweet is directed towards incels themselves and not for example women.

Finally, as can be seen from the small selection of tweets above, the misogynism score seems to capture not only hateful and misogynist content but rather incel slang in general and thus unexpectedly high misogynist scores are attributed to relatively innocent tweets that include incel slang.

It is crucially important to remember those limitations when having a look at the results of the analysis. However, the author decided to proceed as overall the benefits, especially the amount of data that can be considered, trump the potential limitations.

5. Results

This chapter presents the results of the analyses which rely on OLS multiple regression by focusing initially on the first and then on the second hypothesis. Table 7 displays the results of the four models constructed to test the first hypothesis: *The more toxic a Tweet, the more often the Tweet is retweeted, within the incel community on Twitter.*

Regression Results H1

	Dependent variable: Retweet Count			
	OLS		*panel linear*	
	(1)	(2)	(3)	(4)
Toxicity Score	-0.044		0.182^{***}	
	(0.043)		(0.059)	
Toxicity Binary		-0.043^{***}		0.041^{***}
		(0.010)		(0.014)
Presence of URLs			0.009	0.013
			(0.019)	(0.019)
Presence of Hashtags			0.209^{***}	0.211^{***}
			(0.026)	(0.026)
Number of Followers			0.177^{***}	0.175^{***}
			(0.026)	(0.026)
Constant	0.216^{***}	0.246^{***}		
	(0.005)	(0.009)		
Observations	18,826	18,826	18,826	18,826
R^2	0.0001	0.001	0.015	0.014
Adjusted R^2	0.00000	0.001	-1.228	-1.229
Residual Std. Error (df = 18824)	0.602	0.602		
F Statistic	1.034 (df = 1; 18824)	18.113^{***} (df = 1; 18824)	30.897^{***} (df = 4; 8324)	30.583^{***} (df = 4; 8324)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 7: Regression Results H1

Overall, the results are mixed. The first model, which includes only the toxicity score, results in a negative and non-significant coefficient. Moreover, the F-statistic, assessing the overall significance of the model is non-significant (p-value=0.3093) and thus the model does not seem fit to explain variation in retweet counts.

The second model focuses on the binary toxicity variable. The F-statistic indicates that the overall model fit is significant ($p < 0.01$). However, the R^2 value is rather low in that the model explains only 0.01% of the variation in the dependent variable. The coefficient for the binary toxicity variable is significant ($p < 0.01$) but against my expectations negative. This means that toxic tweets in comparison to non-toxic tweets are associated with lower retweet counts. This negative relationship is little surprising when considering that the correlation between the toxicity score and the retweet count is slightly negative as well (r= -0.0074). This negative coefficient for the toxicity scores in both versions of the variable disappears when including potential confounders in the model. Model 3 includes not only the toxicity score but also the presence of URLs, the presence of hashtags and the number of followers based on user-fixed effects. The F-statistic indicates that the model overall is significant ($p < 0.01$). The R^2 value is slightly higher and implies that the model accounts for 1.5% of the variation in the dependent variable. According to model 3, the independent variable toxicity score has a positive coefficient which is significant ($p < 0.01$). Accordingly, an increase in the toxicity score is associated with an 18.2% increase in retweet counts. The individual contribution of each of these variables can be seen when looking at the standardised coefficients (Table 8).

Standardised Coefficients Model 3

Toxicity Score	Presence of URLs	Presence of Hashtags	Number of Followers
0.001	0.149	0.127	0.032

Table 8: Standardised Coefficient Model 3

Consequentially, the relative contribution of both the presence of hashtags and the number of followers is considerably higher than the contribution of the toxicity score, which is not to be neglected despite being rather small. Additionally, after checking for homogeneity of variance both graphically and analytically with the help of the Breusch-Pagan test, I decided to include robust standard errors (Table 9) to be sure that the heterogeneity of variance does not lead to biased results through distorted standard errors. Despite the now higher robust standard errors, the significance of the toxicity scores ($p < 0.01$) is confirmed. Additionally, I tested analytically

Robust Standard Errors Model 3:

	Dependent variable:
	Retweet Count
Toxicity Score	0.182***
	(0.064)
Presence of URLs	0.009
	(0.050)
Presence of Hashtags	0.209***
	(0.061)
Number of Followers	0.177***
	(0.050)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Robust Standard Errors Model 3

the assumption of independence of errors. The results of the Durbin Watson test indicate that the assumption can be confirmed.

Finally, the fourth and last model includes the binary toxicity variable as well as the three potential confounders. The F-statistic again indicates that the overall model fit is significant ($p < 0.01$). Moreover, according to the R^2 value, the model explains 1.4% of the variation in the retweet count. The coefficient for the binary toxicity variable is significant ($p < 0.01$) and as expected positive. This is an important finding as the skewness of the toxicity score distribution could have influenced the results. Therefore, it is important that the binary variable confirms the findings of model 3. More concretely, switching the binary toxicity variable from non-toxic to toxic is associated with a 4.1% increase in the retweet count variable. When looking at the standardised coefficients of model 4 (Table 10), the relatively higher contribution of the presence of hashtags and the number of followers can be confirmed. However, as already stated above, the effect of the binary toxicity variable although small, cannot be neglected.

Standardised Coefficients Model 4

Toxicity Binary	Presence of URLs	Presence of Hashtags	Number of Followers
0.009	0.150	0.125	0.049

Table 10: Standardised Coefficients Model 4

Additionally, I again decided to include robust standard errors as the Breusch-Pagan test indicated heterogeneity of variance (Table 11). Consequentially, the higher standard errors lead to a loss of significance as the coefficient for the binary toxicity variable is now only significant at the 10% level while the presence of hashtags and the number of followers remain highly significant. Again, I also tested the assumption of independence of errors, which can be confirmed by the results of the Durbin Watson test. Overall, I found support for my hypothesis, particularly in model 3, which included the potential confounders and user-fixed effects. Furthermore, as I corrected for heterogeneity of variance with robust standard errors, successfully excluded the possibility of perfect multicollinearity between the independent variable and any of the control variables and confirmed the independence of errors, I can generalise my sample findings to the incel population on Twitter. Therefore, I can confirm that more toxic tweets are retweeted more often within the incel community on Twitter.

Robust Standard Errors Model 4:

	Dependent variable:
	Retweet Count
Toxicity Binary	0.041^{*}
	(0.025)
Presence of URLs	0.013
	(0.047)
Presence of Hashtags	0.211^{***}
	(0.061)
Number of Followers	0.175^{***}
	(0.049)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 11: Robust Standard Errors Model 4

Now I'll turn to test the second hypothesis which states: *The higher the percentage of misogynist words in a Tweet, the more often the Tweet is retweeted in the incel community on Twitter.* The regression results are reported in Table 12. Only model 8, the model based on the binary misogynism variable that also controls for potential confounders has a significant coefficient ($p < 0.01$). The F-statistic suggests that the model is overall significant while the R^2 value indicates that the model explains 6.5% of the variation in the dependent variable. According to model 8, switching the binary misogynism variable from non-misogynist to misogynist is associated with a 3.2% increase in the retweet count. When looking at the individual contribution of each of these variables with the help of standardised coefficients (Table 13), it can be seen that the contribution of the binary misogynism variable is relatively small when comparing it with non-content factors. However, it should not be neglected.

Regression Results H2

	Dependent variable: Retweet Count			
	OLS		*panel linear*	
	(5)	(6)	(7)	(8)
Misogynism Score	-0.162		0.027	
	(0.126)		(0.110)	
Misogynism Binary		0.013		0.032***
		(0.013)		(0.011)
Presence of URLs			0.026**	0.027**
			(0.011)	(0.011)
Presence of Hashtags			0.163***	0.163***
			(0.012)	(0.012)
Number of Followers			0.074***	0.074***
			(0.002)	(0.002)
Constant	0.215***	0.212***	-0.235***	-0.239***
	(0.005)	(0.005)	(0.013)	(0.013)
Observations	19,359	19,359	19,359	19,359
R^2	0.0001	0.00005	0.065	0.065
Adjusted R^2	0.00003	-0.00000	0.065	0.065
Residual Std. Error (df = 19357)	0.602	0.602		
F Statistic (df = 1; 19357)	1.650	0.933	1,262.851***	1,270.867***
Note:			*p<0.1; **p<0.05; ***p<0.01	

Table 12: Regression Results H2

As above, I again decided to include robust standard errors as the Breusch-Pagan test indicated heterogeneity of variance (Table 14). The robust standard errors led to a slight loss of significance, however it remained significant (p < 0.05). Again, I also tested the assumption of independence of errors, which can be confirmed by the results of the Durbin Watson test.

Standardised Coefficients Model 8

Misogynism Binary	Presence of URLs	Presence of Hashtags	Number of Followers
0.017	0.019	0.116	0.277

Table 13: Standardised Coefficients Model 8

Robust Standard Errors Model 8:

	Dependent variable:
	Retweet Count
Misogynism Binary	0.032^{**}
	(0.013)
Presence of URLs	0.027
	(0.022)
Presence of Hashtags	0.163^{***}
	(0.021)
Number of Followers	0.074^{***}
	(0.004)
Constant	-0.239^{***}
	(0.017)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 14: Robust Standard Errors Model 8

Overall, there is support in favour of my second hypothesis. Although only one model provided a significant coefficient, model 8 is a particularly important one as it controls for potential confounders and includes user-fixed effects. Similarly, it safeguards against biases caused by skewness of variable distribution which can be problematic in the continuous variable version. Similar to above I corrected for heterogeneity of variance with robust standard errors, successfully excluded the possibility of perfect multicollinearity between the independent variable and any of the control variables and confirmed the independence of errors. Therefore, I can confirm that more misogynist tweets are retweeted more often within the incel community on Twitter.

6. Discussion

In the previous chapter, I tested my two hypotheses claiming that more toxic and more misogynist tweets are retweeted more often within the incel community by using OLS regression. Regarding the first hypothesis I found, when controlling for the potential confounder presence of hashtags and number of followers, that more toxic tweets are according to my expectations retweeted more often. This was true for the continuous and the binary version of the toxicity variable although to a less significant extent in the latter case. More concretely, an increase in the continuous toxicity score was associated with an increase in the retweeting count by 18.2% while switching the binary toxicity variable from 0 to 1 was associated with an increase in the retweeting count by 4.1%. The analysis thus confirms a positive effect of the toxicity score of a tweet on its retweetability and thus aligns with the findings of the information diffusion literature in that more negative or more hateful content spreads further on Twitter (Fan et al., 2016; Jenders et al., 2013; Kim & Yoo, 2012; Naveed et al., 2011). This within-community analysis thus confirms this trend for the incel community on Twitter. This is particularly astonishing as the incel community as such is already a rather toxic environment. Nevertheless, the results show that varying toxicity levels within an already toxic community still lead to variations in the retweetability of tweets.

When switching the focus from the toxicity to the misogynism of tweets, the results of the second analysis confirm the second hypothesis. Consequentially, I found when focusing on the binary misogynism variable and controlling for the potential confounder presence of hashtags, presence of URLs, and number of followers, that more misogynist tweets are according to my expectations retweeted more often. More concretely, switching the binary toxicity variable from 0 to 1 was associated with an increase in the retweeting count by 3.2%.

Nevertheless, the effect is rather small. A potential explanation can be traced to a limited internal validity or more concretely limited construct validity. Taylor (2013, p. 143) claims that construct validity revolves around the question of "whether inferences from test scores are appropriate". Essential to this is whether "scores can be trusted" (ibid.). I argue, however, that the misogynism score does not entirely capture misogynism but rather the usage of incel slang and might thus be conceptionally too broad. The demonstrations in Table 6 reported randomly selected tweets and their respective misogynism scores. As already discussed above some tweets seem to have received unexpectedly high misogynism scores when the tweets are indeed not misogynist but rather employ words associated with incel slang. This substantially biased the results as the misogynism score does not measure what it is supposed to measure. Therefore, the findings have to be treated cautiously. Additionally, as demonstrated above the dictionary is not able to differentiate between hate directed towards others and self-hate. This is the case because the dictionary approach in its nature focuses on matches with words within the dictionary while not being able to account for contextual factors. This however is crucial because a substantial part of the incel ideology centres on self-hate caused by disappointment about not being recognised by the society as neither attractive nor socially successful (for example Daly & Reed, 2022). Moreover, the findings of Kim and Yoo (2012) indicate that tweets that are predominately perceived as sad, are negatively correlated with retweeting behaviour. Consequentially, tweets that wrongfully receive high misogynism scores while in fact revolving around self-hate might crucially bias the results.

Another factor that needs to be considered is Twitter's content moderation in cases of Twitter rule violation. The Twitter rules are dedicated "to ensure all people can participate in the public conversation freely and safely".[11] Accordingly, Twitter defines types of behaviour that "discourage[s] people from expressing themselves, and ultimately diminish[es] the value of global public conversation" (ibid.) and are thus prohibited. These types of behaviour include among others threatening with or glorifying violence, promoting terrorism or violent extremism, targeted harassment or hate based on for example the race, sexual orientation or gender of other people (ibid.). Consequentially, different rule enforcement actions can be directed either against an individual tweet or the responsible user. These actions can have different severity levels, ranging from labelling the tweet as misleading, limiting the tweets visibility, turning off engagement options with the tweets -including the possibility to retweet- and as ultima ration also include the removal of the tweet. If violations are particularly severe or occur repeatedly users can be suspended as well.[12]

The focus of this research overlaps with those types of behaviour that are prohibited according to the Twitter rules. This is problematic because deleted content can substantially differ from not deleted content demonstrated by King et al., (2013). Consequentially, the data collection process might be severely limited to the prior content moderation and subsequent deletion of radical content. This, however, would have crucial implications not only for the data set as such but also for the dependent variable. The retweet count is naturally biased when engagement with a radical tweet is disabled, or the tweet is even deleted and thus cannot be retweeted. The author is unable to assess the extent of the influence of content moderation on

[11] Twitter. Rules and Policies. Retrieved from: https://help.twitter.com/en/rules-and-policies/twitter-rules, checked 09.06.2022.

[12] Twitter. Rule Enforcement Options. Retrieved from: https://help.twitter.com/en/rules-and-policies/enforcement-options, checked 09.06.2022.

the results or the extent to which tweets were deleted before the data collection. Even though the data collection was repeated daily a time frame of roughly 24 hours theoretically made it possible for tweets to be tweeted, flagged, and removed and thus missed out by the data collection. However, to get an idea of the extent of deleted tweets, I rerun the data collection four weeks after its initial termination based on the tweet IDs and found that of the 19,359 initially collected tweets 16,582 were still online. At the same time, this means that 14.34% were removed. There can of course be multiple reasons why those tweets are not online anymore including not only the removal due to content moderation but also trivial aspects such as users reconsidering their tweeting decisions and thus removing tweets or users leaving Twitter and deleting their accounts overall. Nevertheless, 14.34% of removed tweets are a crucial share, especially when paired with the finding that the deleted tweets are significantly more toxic ($p < 0.01$) than the whole data set with a mean of 0.069 compared to 0.052. The same is true for the misogynism score which is on average higher in the removed group (0.012) than in the overall data set (0.010). This finding is significant ($p < 0.01$) as well. Consequentially, the possibility that Twitter's content moderation influenced my results cannot be excluded. While repeating the data collection process every day possibly limited this effect on missing out on tweets. The same cannot be said about the effect on the dependent variable. Hence, even if tweets originally might have been included in the dataset their respective retweets count might have grown over subsequent days which, however, was not possible due to content removal. Perhaps the incel community also is aware of the danger encompassed by content moderation and thus found a way to bypass tweet removal. Hence, the effect would be smaller than anticipated. But either way purposely using less radical language ultimately also influences the results in a way that the toxicity and misogynism scores are kept within limits deliberately. Therefore, the effect of content moderation necessarily must be considered when analysing the results.

Another important aspect is that the study design of this analysis is a within-community analysis which thus focuses exclusively on an already radical environment. Consequentially, the findings might underestimate the effect of toxicity or misogynism on retweeting behaviour as the variation of both is possibly lower within the incel community compared to other twitter environments. An overall high level of toxicity or misogynism might thus limit the extent of the influence. Therefore, what seems like a comparably weak result is in fact a considerable outcome.

This also relates to the question of the generalisability of the findings. As stated above the incel community is drastically different from the overall twitter environment. The findings are thus naturally limited to the incel community as such. This is particularly the case as the incel-specific echo-chamber effect, that drives polarisation and radicalisation within one community does not affect the overall twitter environment. Hence, there is no reason to expect a relationship between toxicity or misogynism and the retweeting behaviour of general Twitter users.

Moreover, retweeting behaviour can have various different motivations (Boyd et al., 2010). This gets even more complicated when leaving the incel community. Retweeting toxic or misogynist tweets can apart from voicing agreement also be used to flag problematic content or to express outrage. Consequentially retweeting is not necessarily an intentional tool to spread ideology.

Another important aspect is what Daly and Reed (2022) describe as "shitposting" as it also crucially biases both the toxicity and the misogynism score and thus eventually also the results. They argue that incels practice "shitposting" intentionally to provoke and to shock while this does thus not properly represent incel ideology where "the majority of incels do not commit acts of physical violence" (ibid., p. 17). Therefore, Daly and Reed (20022) argue that incel ideology seems far more radical than it is. Similarly, Speckhard et al. (2021) argue that analysis based on online content can produce biased impressions of incel radicalness. They claim that online content "may represent exaggeration amidst group polarization, rather than an actual representation of true beliefs, attitudes and behaviors" (ibid., p. 92). However, while it might be true that incels feel particularly compelled to exacerbate and express their opinion in the most shocking way online, I argue that this radical way of expression nevertheless has important implications as it influences and pushes "the limits and forms of the sayable" (Foucault & Sheridan, 2012). Consequentially, whether intended or not, the discourse radicalises further while misogynist threats become normalised.

A few limitations should be considered at this point. I think it is natural that keyword-based Twitter data collection can only capture an extract of the whole discourse of a community. At the same time, the data set might be biased as other topics and communities are accidentally included. Therefore, Firdaus et al., (2018) suggest an alternative data collection approach to explore large connected networks, namely snowball sampling. In this approach interactions – mentioning or retweeting – of accounts that are clearly identified as belonging to the incel community are traced until a sufficient data set is generated. Therefore, the data collection process is guided by the users rather than by the content following the expectation to receive a data set which represents the community. Moreover, a lot of information is lost by not accounting for the content of pictures or URLs. However, this content rather than its description might be deciding the question of whether a tweet is retweeted or not. Further research could thus focus on including this information for example with the help of visual topic modelling (Střítecký & Špelda, 2017). Finally, as discussed above in detail, the measurement for misogynism certainly needs improvement as the dictionary approach seems to be insufficient. Consequentially, future research could focus on verifying my findings in regard to the misogynism of tweets with an improved measurement through for example super-vised learning methods.

Keeping these limitations in mind, I argue that the contribution of this thesis is not to be neglected. The results of my analysis offer a first cautious trend about the retweeting behaviour within the incel community, a so-far unexplored topic. Future research can build on my findings and gain deeper insights into the diffusion of radical context within the incel community on Twitter.

This analysis found that more toxic content spreads further within the incel community. This has crucial implications for the radicalisation potential of the incel community on Twitter, especially when paired with the findings of Speckhard et al. (2021) who found that participation in online incel forums tends to further strengthen and radicalise opinions of already violent and misogynist users through a process of reaffirmation and validation with converging opinions. Frequent exposure to the most toxic content can not only lead to a critical normalisation of violence-inciting attitudes within the incel echo chamber but eventually also influence individuals who just entered the community and got in touch with incel ideology for the first time. Twitter has a particular role in that it does not have pre-defined communities and content can thus easily spread beyond the incel community. At the same time, Twitter is a public and open platform. The spreading of radical incel content is thus worrying as it indicates that open misogynism and violent attitudes left the shadows and can be openly displayed without consequences. Ultimately, constant reaffirmation of one's own darkest beliefs might accelerate the radicalisation of others within the incel community. Therefore, online discourses cannot be regarded as a niche phenomenon but have to be understood as what they are: breeding grounds for desperate and violence-seeking individuals.

7. Conclusion

This thesis investigated the research question: *To what extent do more radical Tweets diffuse further within the incel community?* To analyse this question, I gathered a total of 52,927 tweets over the course of seven weeks based on keywords prominent within the incel community. I investigated two types of radical content – toxic and misogynist tweets. The former was defined with the Perspective API, a machine learning algorithm used for content moderation and determining the severe toxicity of a tweet. The misogynism of tweets was determined with the help of a dictionary approach based on a dictionary developed by Farrell et al. (2019) where matches were normalised with the overall word count of the respective tweet. As the method used to answer the research question, the author employed multiple models of ordinary least square regression with and without controlling for potential confounders and in a continuous and binary variable version. Moreover, user-fixed effects were integrated into the model. The results supported the hypothesis that toxic tweets are retweeted more often within the incel community on Twitter. More concretely, an increase in the toxicity score of a tweet was found to be associated with an 18.2% increase in retweets. At the same time, misogynist tweets are retweeted 3.2% more often than non-misogynist tweets.

Consequentially, my results indicate that more toxic and misogynist tweets diffuse further within the incel community. This is a significant finding as it has important implications for the radicalisation potential of the incel community on Twitter. As more radical content tends to spread further, more users are exposed to radical ideas. Ultimately, exposure to radical incel ideology and constant reaffirmation of hopeless injustices can crucially amplify the radicalisation of others. Although Incel terrorists in the past acted alone when committing their attacks, they were part of a wider online community from which they received not only support but also an encouragement to make the decision to take part in the incel rebellion. The diffusion of radical incel content on Twitter can thus not be underestimated. Ensuring free speech on Twitter, as currently envisioned by Elon Musk, therefore does inherit certain risks. Twitter's current reluctance (Gorwa et al., 2020) when it comes to content moderation might be further reduced after a potential take-over from Musk, thus eventually giving more space for radical ideas to flourish and spread.

This thesis contributed to the scholarly debate by finding that more toxic and more misogynist content diffuses further in the incel community on Twitter. By doing so I contributed to closing the research gap about the radicalisation potential of the incel community on Twitter while shedding light on the danger of violent misogynism and incel radicalisation. It thus provides a first trend on which further research can build when exploring the radicalisation potential of the incel community on Twitter.

References

Adewale Obadimu, Mead, E., Hussain, M. N., & Nitin Agarwal. (2019). *Identifying Toxicity Within YouTube Video Comment Text Data*. https://doi.org/10.13140/RG.2.2.15254.19522

Aghazadeh, S. A., Burns, A., Chu, J., Feigenblatt, H., Laribee, E., Maynard, L., Meyers, A. L. M., O'Brien, J. L., & Rufus, L. (2018). GamerGate: A Case Study in Online Harassment. In J. Golbeck (Ed.), *Online Harassment* (pp. 179–207). Springer International Publishing. https://doi.org/10.1007/978-3-319-78583-7_8

Agius, C., Edney-Browne, A., Nicholas, L., & Cook, K. (2021). Anti-feminism, gender and the far-right gap in C/PVE measures. *Critical Studies on Terrorism*, 1–25. https://doi.org/10.1080/17539153.2021.1967299

Alizadeh, M., Gilardi, F., Hoes, E., Klueser, J., Kubli, M., & Marchal, N. (2022). *Content moderation as a political issue: The Twitter discourse around Trump's ban*.

Allely, C. S., & Faccini, L. (2017). "Path to intended violence" model to understand mass violence in the case of Elliot Rodger. *Aggression and Violent Behavior*, *37*, 201–209. https://doi.org/10.1016/j.avb.2017.09.005

Anderson, B. R. O. (2016). *Imagined communities: Reflections on the origin and spread of nationalism* (Revised edition). Verso.

Anzovino, M., Fersini, E., & Rosso, P. (2018). Automatic Identification and Classification of Misogynistic Language on Twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, & F. Meziane (Eds.), *Natural Language Processing and Information Systems* (Vol. 10859, pp. 57–64). Springer International Publishing. https://doi.org/10.1007/978-3-319-91947-8_6

Baele, S. J., Brace, L., & Coan, T. G. (2021). From “Incel” to “Saint”: Analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and Political Violence*, *33*(8), 1667–1691. https://doi.org/10.1080/09546553.2019.1638256

Bartlett, J., Norrie, R., Patel, S., Rumpel, R., & Wibberley, S. (2014). *Misogyny on Twitter*. Demos. https://demos.co.uk/

BBC. (2016, February 5). Twitter suspends 125,000 ‘terrorism’ accounts. *BBC*. https://www.bbc.com/news/world-us-canada-35505996

Berger, J. M. (2018). *Extremism*. The MIT Press.

Borum, R. (2003). Understanding the Terrorist Mind-Set. *FBI Law Enforcement Bulletin*.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *2010 43rd Hawaii International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2010.412

Brace, L. (2021). A short introduction to the involuntary celibate sub-culture. *Crest Article*. https://crestresearch.ac.uk/projects/tracking-online-contagion-incel/

Brooks, R. C., Russo-Batterham, D., & Blake, K. R. (2022). Incel Activity on Social Media Linked to Local Mating Ecology. *Psychological Science*, *33*(2), 249–258. https://doi.org/10.1177/09567976211036065

Brüderl, J., & Ludwig, V. (2014). Fixed-Effects Panel Regression. In H. Best & C. Wolf, *The SAGE Handbook of Regression Analysis and Causal Inference* (pp. 327–358). SAGE Publications Ltd. https://doi.org/10.4135/9781446288146.n15

Clancy, T., Addison, B., Pavlov, O., & Saeed, K. (2021). Contingencies of Violent Radicalization: The Terror Contagion Simulation. *Systems*, *9*(4), 90. https://doi.org/10.3390/systems9040090

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data:

Political Homophily on Twitter. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Comarela, G., Crovella, M., Almeida, V., & Benevenuto, F. (2012). Understanding factors that affect response rates in twitter. *Proceedings of the 23rd ACM Conference on Hypertext and Social Media - HT '12*, 123. https://doi.org/10.1145/2309996.2310017

Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic Masculinity: Rethinking the Concept. *Gender & Society*, *19*(6), 829–859. https://doi.org/10.1177/0891243205278639

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771.

Daly, S. E., & Reed, S. M. (2022). "I Think Most of Society Hates Us": A Qualitative Thematic Analysis of Interviews with Incels. *Sex Roles*, *86*(1–2), 14–33. https://doi.org/10.1007/s11199-021-01250-5

Fan, R., Xu, K., & Zhao, J. (2016). *Higher contagion and weaker ties mean anger spreads faster than joy in social media*. https://doi.org/10.48550/ARXIV.1608.03656

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring Misogyny across the Manosphere in Reddit. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 87–96. https://doi.org/10.1145/3292522.3326045

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Filippo, M., Fulper, R. S., Ferrara, E. L., Ahn, Y., Flammini, A., Lewis, B., & Rowe, K. K. (2015). *Misogynistic Language on Twitter and Sexual Violence*.

Firdaus, S. N., Ding, C., & Sadeghian, A. (2018). Retweet: A popular information diffusion mechanism – A survey paper. *Online Social Networks and Media*, *6*, 26–40. https://doi.org/10.1016/j.osnem.2018.04.001

Foucault, M., & Sheridan, A. (2012). *Discipline and punish: The birth of the prison*. Vintage. http://0-lib.myilibrary.com.catalogue.libraries.london.ac.uk?id=435863

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4743–4752. https://doi.org/10.3233/JIFS-179023

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press. http://www.SLQ.eblib.com.au/patron/FullRecord.aspx?p=288457

Ging, D. (2019). Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, *22*(4), 638–657. https://doi.org/10.1177/1097184X17706401

Gold, N., & Department of Computer Science at UCL. (2020). *Using Twitter Data in Research Guidance for Researchers and Ethics Reviewers*. https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Greene, W. H. (2012). *Econometric analysis* (7th ed). Prentice Hall.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gruzd, A., Wellman, B., & Takhteyev, Y. (2011). Imagining Twitter as an Imagined Community. *American Behavioral Scientist*, *55*(10), 1294–1318. https://doi.org/10.1177/0002764211409378

Hafez, M., & Mullins, C. (2015). The Radicalization Puzzle: A Theoretical Synthesis of Empirical Approaches to Homegrown Extremism. *Studies in Conflict & Terrorism*, *38*(11), 958–975. https://doi.org/10.1080/1057610X.2015.1051375

Hajarian, M., & Khanbabaloo, Z. (2021). Toward Stopping Incel Rebellion: Detecting Incels in Social Media Using Sentiment Analysis. *2021 7th International Conference on Web Research (ICWR)*, 169–174. https://doi.org/10.1109/ICWR51868.2021.9443027

Hamm, M. S., & Spaaij, R. F. J. (2017). *The age of lone wolf terrorism*. Columbia University Press.

Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, *91*, 80–93. https://doi.org/10.1016/j.pragma.2015.11.005

Hermansson, P., Lawrence, D., Mulhall, J., & Murdoch, S. (2020). *The international alt-right: Fascism for the 21st century?* Routledge.

Hoffman, B., Ware, J., & Shapiro, E. (2020). Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism*, *43*(7), 565–587. https://doi.org/10.1080/1057610X.2020.1751459

Horgan, J. (2008). From Profiles to Pathways and Roots to Routes: Perspectives from Psychology on Radicalization into Terrorism. *The ANNALS of the American Academy of Political and Social Science*, *618*(1), 80–94. https://doi.org/10.1177/0002716208317539

Horta Ribeiro, M., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., & Zannettou, S. (2021). The Evolution of the Manosphere across the Web. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 196–207.

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. https://doi.org/10.1145/3476057

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). *Deceiving Google's Perspective API Built for Detecting Toxic Comments*. https://doi.org/10.48550/ARXIV.1702.08138

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, *7*(2), 240–268. https://doi.org/10.1075/jlac.00026.jak

Jane, E. A. (2014). 'Back to the kitchen, cunt': Speaking the unspeakable about online misogyny. *Continuum*, *28*(4), 558–570. https://doi.org/10.1080/10304312.2014.924479

Jenders, M., Kasneci, G., & Naumann, F. (2013). Analyzing and predicting viral tweets. *Proceedings of the 22nd International Conference on World Wide Web*, 657–664. https://doi.org/10.1145/2487788.2488017

Kantrowitz, A. (2019, July 23). The Man Who Built The Retweet: "We Handed A Loaded Weapon To 4-Year-Olds". *BuzzFeed.News*. https://www.buzzfeednews.com/article/alexkantrowitz/how-the-retweet-ruined-the-internet

Kearney, M. W., Heiss, A., & Briatte, F. (2020). *Rtweet*. https://cran.r-project.org/web/packages/rtweet/rtweet.pdf

Kim, J., & Yoo, J. (2012). Role of Sentiment in Message Propagation: Reply vs. Retweet Behavior in Political Communication. *2012 International Conference on Social Informatics*, 131–136. https://doi.org/10.1109/SocialInformatics.2012.33

King, G., Pan, J., & Roberts, M. E. (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, *107*(2), 326–343. https://doi.org/10.1017/S0003055413000014

Kleinman, Z. (2022, April 25). Twitter: Why Elon Musk has been so keen on taking control. *BBC*. https://www.bbc.com/news/technology-61222793

Langman, P. (2016). Elliot Rodger: An Analysis. *The Journal of Campus Behavioral Intervention*.

Makkar, S., & Chakraborty, T. (2020). *Hate speech diffusion in twitter social media*. IIIT-Delhi.

Mantilla, K. (2013). Gendertrolling: Misogyny Adapts to New Media. *Feminist Studies*, *39*(2), 563–570.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*, 173–182. https://doi.org/10.1145/3292522.3326034

Milmo, D. (2022, April 16). Banned from Twitter: Accounts that may be reprieved after Musk takeover. *Guardian*. https://www.theguardian.com/technology/2022/apr/26/banned-from-twitter-accounts-reprieve-elon-musk

Minassian, A. (2018, April 23). *Electronically recorded Interview of Alek Minassian by Detective Robert Thomas (3917) of the Sex Crimes Unit Polygraph Unit on Monday, April 23, 2018, at 22:46 Hours* [Personal communication].

Moghaddam, F. M. (2005). The Staircase to Terrorism: A Psychological Exploration. *American Psychologist*, *60*(2), 161–169. https://doi.org/10.1037/0003-066X.60.2.161

Munn, L. (2020). Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), 53. https://doi.org/10.1057/s41599-020-00550-7

Murthy, D. (2012). Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology*, *46*(6), 1059–1073. https://doi.org/10.1177/0038038511422553

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. *Proceedings of the 3rd International Web Science Conference on - WebSci '11*, 1–7. https://doi.org/10.1145/2527031.2527052

Neumann, P. R. (2013). The trouble with radicalization. *International Affairs*, *89*(4), 873–893. https://doi.org/10.1111/1468-2346.12049

O'Malley, R. L., Holt, K., & Holt, T. J. (2022). An Exploration of the Involuntary Celibate (Incel) Subculture Online. *Journal of Interpersonal Violence*, *37*(7–8), NP4981–NP5008. https://doi.org/10.1177/0886260520959625

Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). 'How over is it?' Understanding the Incel Community on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://doi.org/10.1145/3479556

Pariser, E. (2014). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin Books. http://rbdigital.oneclickdigital.com

Perspective API. (2018). https://perspectiveapi.com/

Poland, B. (2016). *Haters: Harassment, abuse, and violence online*. Potomac Books, an imprint of the University of Nebraska Press.

Powell, A., & Henry, N. (2017). Introduction. In A. Powell & N. Henry, *Sexual Violence in a Digital Age* (pp. 1–20). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-58047-4_1

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 557–568.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Rodger, E. (2014). *My Twisted World: The Story of Elliot Rodger*.

Rouda, B. (2020). *'I'd kill for a girl like that': The Black Pill and the Incel Uprising*.

Sang, Y., & Stanton, J. (2020). Analyzing Hate Speech with Incel-Hunters' Critiques. *International Conference on Social Media and Society*, 5–13. https://doi.org/10.1145/3400806.3400808

Scaptura, M. N., & Boyle, K. M. (2020). Masculinity Threat, "Incel" Traits, and Violent Fantasies Among Heterosexual Men in the United States. *Feminist Criminology*, *15*(3), 278–298. https://doi.org/10.1177/1557085119896415

Silber, M., & Bhatt, A. (2007). *Radicalization in the West: The Homegrown Threat*.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). *Analyzing the Targets of Hate in Online Social Media*. https://doi.org/10.48550/ARXIV.1603.07709

Speckhard, A., Ellenberg, M., Morton, J., & Ash, A. (2021). Involuntary Celibates' Experiences of and Grievance over Sexual Exclusion and the Potential Threat of Violence Among Those Active in an Online Incel Forum. *Journal of Strategic Security*, *14*(2), 89–121. https://doi.org/10.5038/1944-0472.14.2.1910

Střítecký, V., & Špelda, P. (2017). Establishing the Complexity of the Islamic State's Visual Propaganda. *Central European Journal of International & Security Studies*, *11*(4).

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *2010 IEEE Second International Conference on Social Computing*, 177–184. https://doi.org/10.1109/SocialCom.2010.33

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, *10*(2), 175–195. https://doi.org/10.1111/1467-9760.00148

Taylor, C. S. (2013). *Validity and validation*. Oxford University Press.

Terren, L., Open University of Catalonia, Borge-Bravo, R., & Open University of Catalonia. (2021). Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research*, *9*, 99–118. https://doi.org/10.12840/ISSN.2255-4165.028

Twitter API. (2020). https://developer.twitter.com/en/products/twitter-api

Van Valkenburgh, S. P. (2021). Digesting the Red Pill: Masculinity and Neoliberalism in the Manosphere. *Men and Masculinities*, *24*(1), 84–103. https://doi.org/10.1177/1097184X18816118

Votta, F. (2021). *PeRspective*. https://cran.r-project.org/web/packages/peRspective/peRspective.pdf

White, S. G. (2017). Case study: The Isla Vista campus community mass murder. *Journal of Threat Assessment and Management*, *4*(1), 20–47. https://doi.org/10.1037/tam0000078

Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and Characterizing Hate Speech on News Websites. *12th ACM Conference on Web Science*, 125–134. https://doi.org/10.1145/3394231.3397902

Appendix

Appendix I

CHARLES UNIVERSITY

Universität Konstanz

DECLARATION

First name: Mia	Last name: Nahrgang
Student ID: 1066432 (Konstanz) & 40955907 (CU)	E-Mail address: Mia.Nahrgang@web.de

I hereby declare that the attached master thesis (title / supervisor):

From Supreme Gentlemen to Incel Rebellion – Analysing the Radicalisation Potential of the Incel Community on Twitter

Prof. Nils B. Weidmann (Konstanz) & doc. PhDr. Vít Střítecký (Charles University)

on the following topic: Diffusion of radical Content on Twitter within the Incel Community

is the result of my own, independent work. I have not used any aids or sources other than those I have referenced in the document.

For contributions and quotations from the works of other people (whether distributed electronically or in hardcopy), I have identified each of them with a reference to the source or the secondary literature. Failure to do so constitutes plagiarism. I will also submit the term paper electronically to the lecturer. Furthermore, I declare that the above-mentioned work has not been otherwise submitted as a term paper.

I am aware that papers that turn out to be plagiarized will be graded "insufficient" (5,0). Every suspected case of plagiarism will be submitted to the examination board, which will decide on further sanctions. The legal basis for this procedure can be found in the examination regulations. These rules also apply to students that study Politics and Public Administration as a minor subject.

Date: 10.07.2022

Signature of the student

Appendix II: Markdown R-Script

Preparations: Used Libraries and Font Type

```r
library(rtweet)
library(tidyverse)
library(peRspective)
library(ggpubr)
library(quanteda)
library(quanteda.textplots)
library(doBy)
library(eeptools)
library(stargazer)
library(plm)
library(sandwich)
library(lmtest)
library(corrplot)
library(QuantPsyc)
library(psych)
library(car)

windowsFonts(Times=windowsFont("Times New Roman"))
```

0. Collecting, Loading and Preparing Twitter Data

```r
#collecting data
#DON'T RUN
api_key <- "XXX"
api_secret_key <-"YYY"
access_token <-"ZZZ"
access_token_secret <- "XYZ"

token <- create_token (app = "ABC", consumer_key = api_key,consumer_secret = api_secret_key, access_token = access_token, access_secret = access_token_secret)

#process repeated every day (Example)
Data_08_05_22<- search_tweets(q="AWLT OR betamale OR betafags OR betauprising OR blackpill OR
                           bluepill OR femoid OR goingER OR incelrevolution
OR mgtow OR redpill OR trueincel",n=18000, include_rts = F)


#save data

saveRDS(Data_08_05_22,"08_05_22")


#create one data frame and remove duplicates

duplicate_detection <-rbind(Data_21_03_22,Data_22_03_22,Data_23_03_22,Data_
24_03_22,Data_25_03_22,Data_26_03_22,Data_27_03_22,Data_28_03_22,Data_29_03
_22,Data_30_03_22,Data_31_03_22,Data_01_04_22,Data_02_04_22,Data_03_04_22,D
ata_04_04_22,Data_05_04_22,Data_06_04_22,Data_07_04_22,Data_08_04_22,Data_0
9_04_22,Data_10_04_22,Data_11_04_22,Data_12_04_22,Data_13_04_22,Data_14_04_
22,Data_15_04_22,Data_16_04_22,Data_17_04_22,Data_18_04_22,Data_19_04_22,Da
ta_20_04_22,Data_21_04_22,Data_22_04_22,Data_23_04_22,Data_24_04_22,Data_25
_04_22,Data_26_04_22,Data_27_04_22,Data_28_04_22,Data_29_04_22,Data_30_04_2
2,Data_01_05_22,Data_02_05_22,Data_03_05_22,Data_04_05_22,Data_05_05_22,Dat
a_06_05_22)


duplicate_reduction <-duplicate_detection %>% group_by(status_id) %>% slice
(which.max(retweet_count))


saveRDS(duplicate_reduction,"sample")

#instead load the data

sample <-readRDS("sample") # 52,927

sample<-subset(sample, lang=="en") # 24,605

sample<- sample %>% filter(user_id != "1414609594849038336" & user_id !="14
40525601211699207"& user_id !="1470052170422788099" & user_id !="1469662887
199260673" & user_id !="1422190368855109636" & user_id !="1109931731036106755"& user_id !="2598894628") #20,795

sample <-subset(sample,display_text_width >19) #19359
```

1. Descriptive Data

1.1 Distribution of Tweets over Time

```
sample$day <-format(as.Date(sample$created_at,format="%y-%m-%d %h:%m:%s"),f
ormat="%y-%m-%d")

sample$day <-as.Date(sample$day,format="%y-%m-%d")


table(sample$day)
```

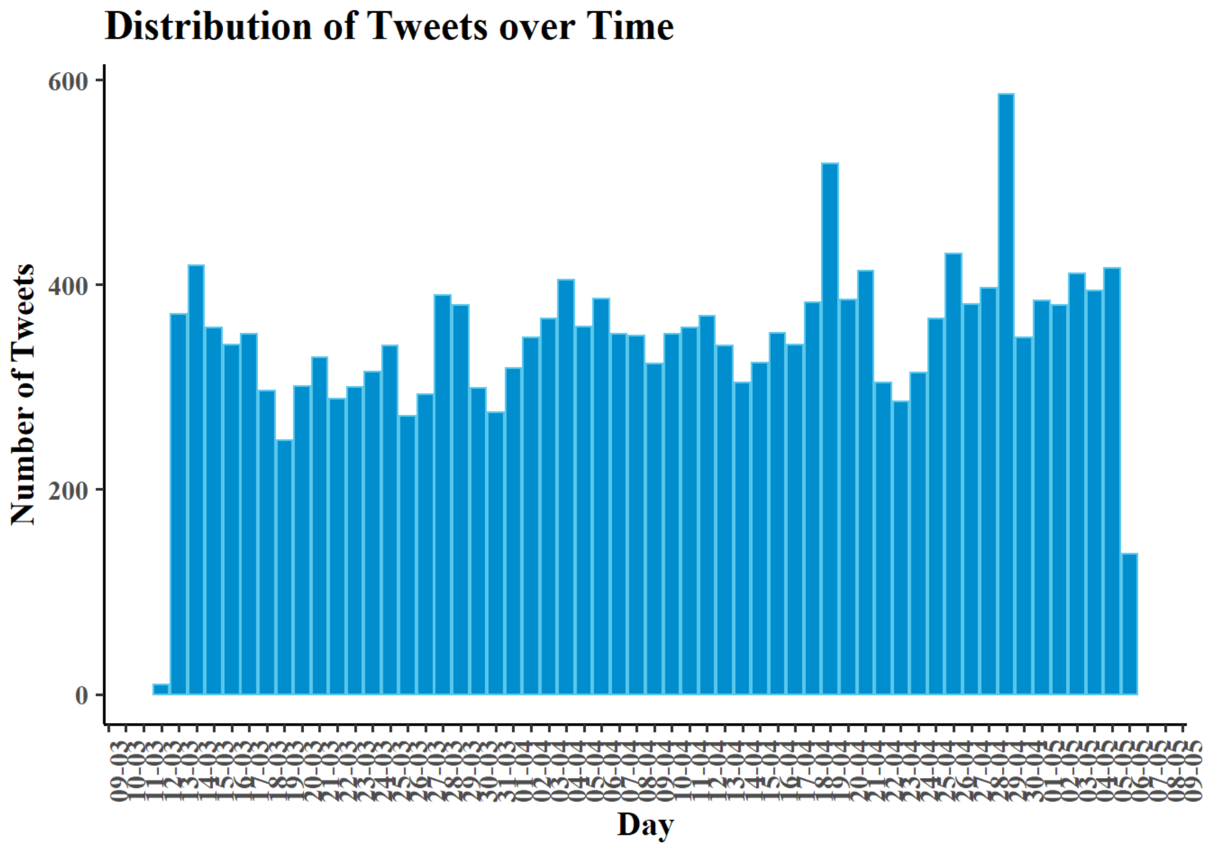
```
##

## 2022-03-12 2022-03-13 2022-03-14 2022-03-15 2022-03-16 2022-03-17 2022-0
3-18

##          9        371        419        358        341        352
296
```

```
## 2022-03-19 2022-03-20 2022-03-21 2022-03-22 2022-03-23 2022-03-24 2022-0
3-25

##        248        301        329        288        300        315
340

## 2022-03-26 2022-03-27 2022-03-28 2022-03-29 2022-03-30 2022-03-31 2022-0
4-01

##        272        293        390        380        299        275
318

## 2022-04-02 2022-04-03 2022-04-04 2022-04-05 2022-04-06 2022-04-07 2022-0
4-08

##        348        367        405        359        386        352
350

## 2022-04-09 2022-04-10 2022-04-11 2022-04-12 2022-04-13 2022-04-14 2022-0
4-15

##        323        352        358        369        340        304
324

## 2022-04-16 2022-04-17 2022-04-18 2022-04-19 2022-04-20 2022-04-21 2022-0
4-22

##        353        341        383        518        385        413
304

## 2022-04-23 2022-04-24 2022-04-25 2022-04-26 2022-04-27 2022-04-28 2022-0
4-29

##        286        314        367        430        381        397
586

## 2022-04-30 2022-05-01 2022-05-02 2022-05-03 2022-05-04 2022-05-05 2022-0
5-06

##        348        384        380        411        394        416
137
```

```
day_count<-summaryBy(text~day,sample,FUN=length)

summary(day_count$text.length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

##     9.0   311.5   352.0   345.7   383.2   586.0
```

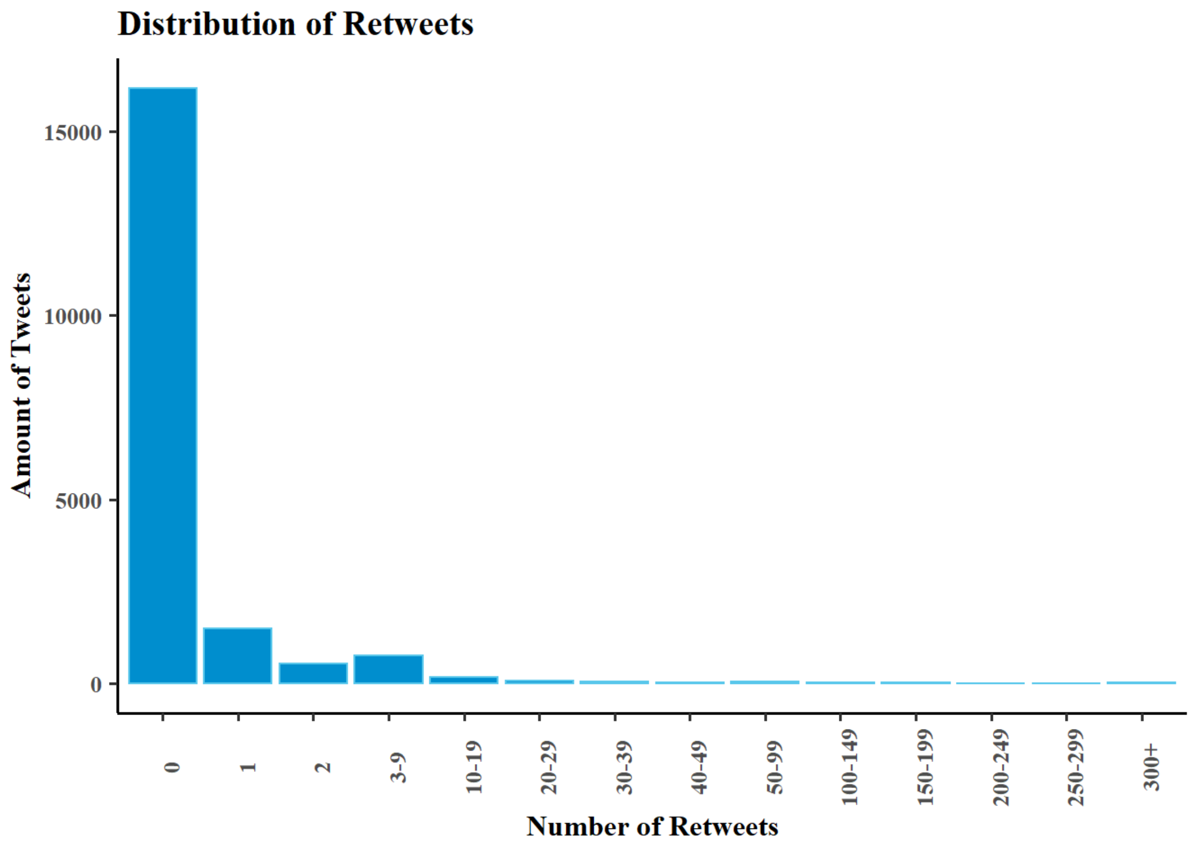
```
tweet_distribution_time <-ggplot(data=sample, aes(x=day))

tweet_distribution_time +geom_bar(color = "#59C7EB", fill ="#008ECE")+theme
_classic(base_size=14)+ylab("Number of Tweets")+xlab("Day") +ggtitle("Distr
ibution of Tweets over Time")+theme(axis.text.x = element_text(angle = 90))
+theme(text=element_text(family="Times", face="bold", size=14))+ scale_x_da
te(date_breaks = 'day', date_labels = '%d-%m')
```



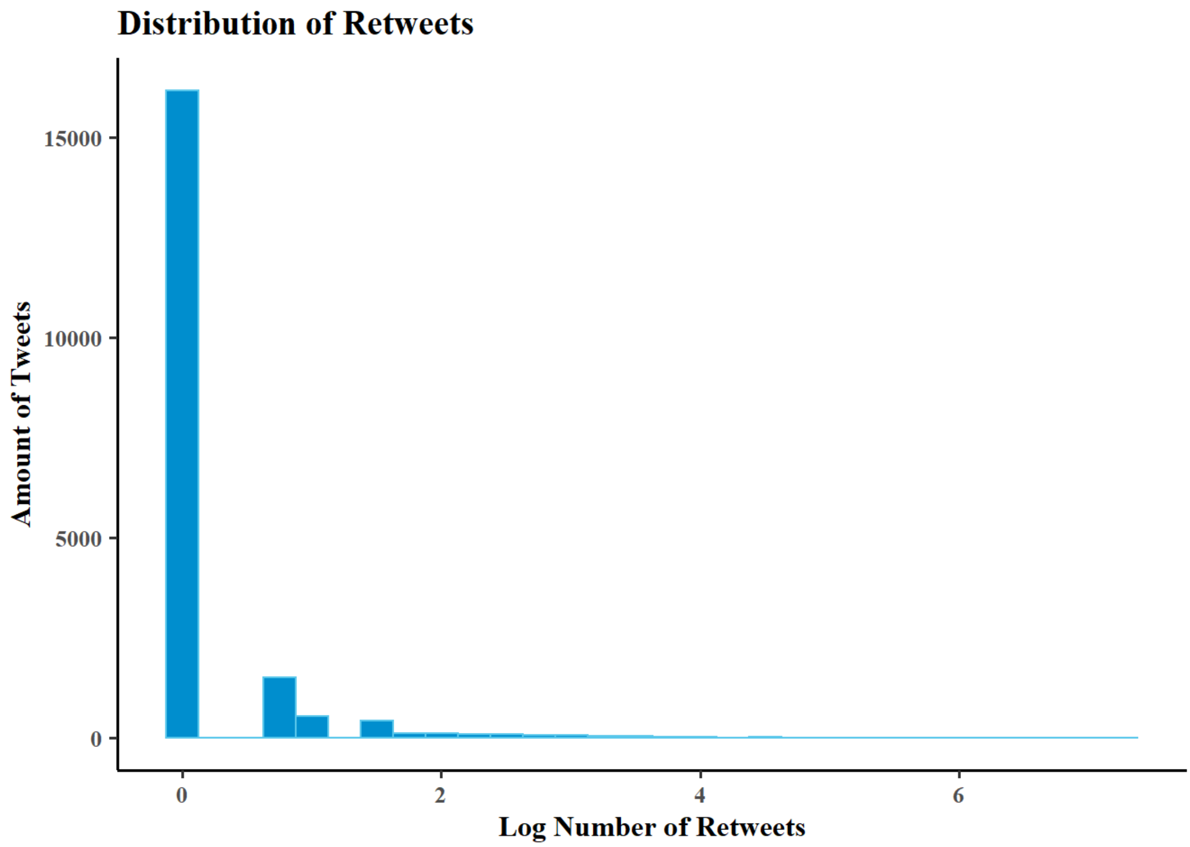
1.2 Distribution of Retweets

```
recount<-summaryBy(text~retweet_count, sample, FUN = length)


sample$retweet_count_50 <- NA

sample$retweet_count_50[sample$retweet_count == 0] <- "0"

sample$retweet_count_50[sample$retweet_count > 0 & sample$retweet_count <=
1] <- "1"

sample$retweet_count_50[sample$retweet_count > 1 & sample$retweet_count <=
2] <- "2"

sample$retweet_count_50[sample$retweet_count < 10 &sample$retweet_count >2]
<- "3-9"

sample$retweet_count_50[sample$retweet_count >=10 & sample$retweet_count <
20] <- "10-19"

sample$retweet_count_50[sample$retweet_count >=20 & sample$retweet_count <
30] <- "20-29"

sample$retweet_count_50[sample$retweet_count >=30 & sample$retweet_count <
40] <- "30-39"

sample$retweet_count_50[sample$retweet_count >=40 & sample$retweet_count <
50] <- "40-49"

sample$retweet_count_50[sample$retweet_count >=50 & sample$retweet_count <
100] <- "50-99"

sample$retweet_count_50[sample$retweet_count >=100 & sample$retweet_count <
150] <- "100-149"
```

```
sample$retweet_count_50[sample$retweet_count >=150 & sample$retweet_count < 200] <- "150-199"

sample$retweet_count_50[sample$retweet_count >=200 & sample$retweet_count < 250] <- "200-249"

sample$retweet_count_50[sample$retweet_count >=250 & sample$retweet_count < 300] <- "250-299"

sample$retweet_count_50[sample$retweet_count >= 300] <- "300+"



retweet_distribution <-ggplot(data=sample, aes(x=retweet_count_50))

positions <- c("0","1","2","3-9","10-19", "20-29","30-39","40-49","50-99","100-149","150-199","200-249","250-299","300+")

retweet_distribution +geom_bar(color = "#59C7EB", fill ="#008ECE")+theme_classic(base_size=14)+xlab("Number of Retweets")+ylab("Amount of Tweets") +ggtitle("Distribution of Retweets")+theme(axis.text.x = element_text(angle = 90))+theme(text=element_text(family="Times", face="bold", size=12))+ scale_x_discrete(limits = positions)
```



```
sample$retweet_count_log <-log1p(sample$retweet_count)

retweet_log_distribution <-ggplot(data=sample, aes(x=retweet_count_log))

retweet_log_distribution + geom_histogram(color = "#59C7EB", fill ="#008ECE")+theme_classic(base_size=14)+xlab("Log Number of Retweets")+ylab("Amount of Tweets") +ggtitle("Distribution of Retweets")+theme(text=element_text(family="Times", face="bold", size=12))
```

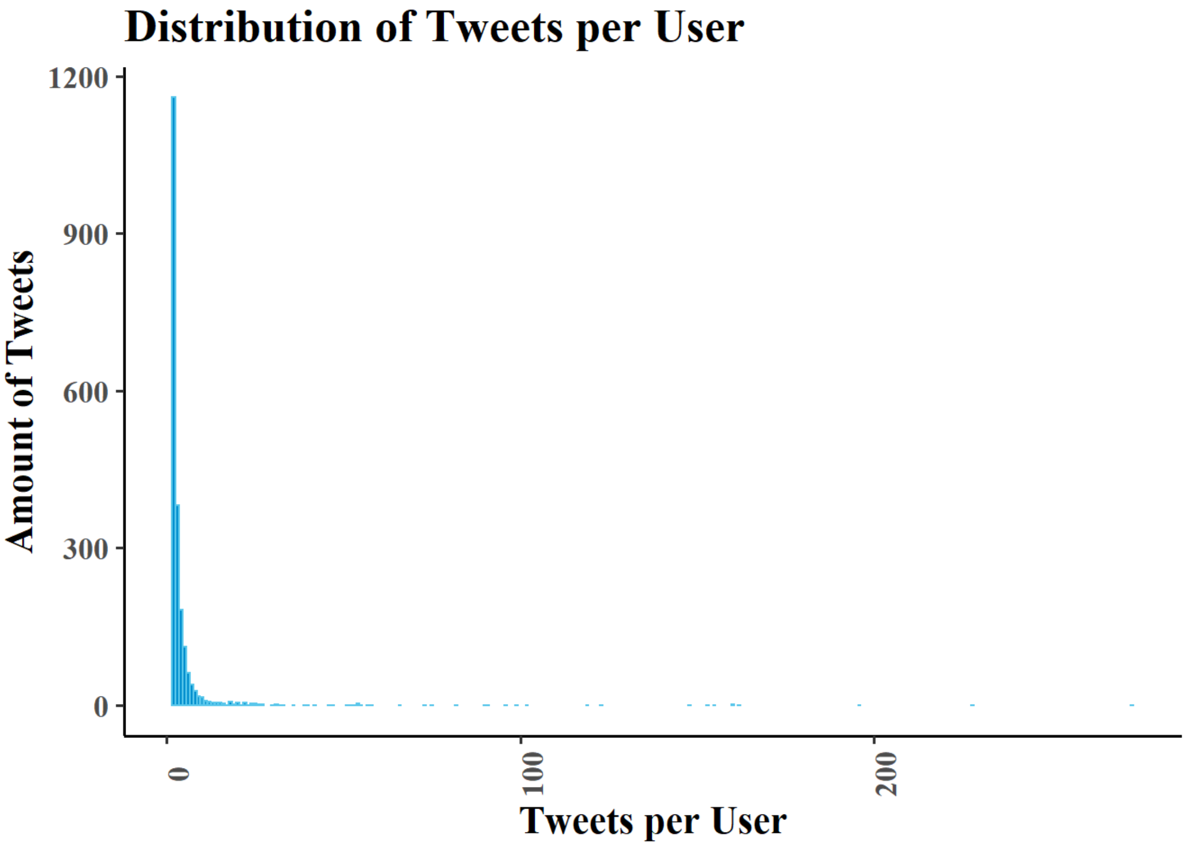


1.3 Distribution of Users

```
user_count<-summaryBy(text~user_id,sample,FUN=length) #10 789 users
mean(user_count$text.length)
## [1] 1.794328
#frequency Table
table(user_count$text.length)
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14
15   16
## 8681 1159  382  182  112   62   40   28   18   16    9    7    6    5
6    3
##   17   18   19   20   21   22   23   24   25   26   27   30   31   32
33   36
##    1    7    2    5    1    5    1    3    4    2    2    1    2    1
1    1
##   39   40   42   46   47   51   52   53   54   55   57   58   66   73
75   82
##    1    1    1    1    1    1    1    1    3    1    1    1    1    1
1    1
##   90   91   96   99  102  119  123  148  153  155  160  162  196  228  2
73
```

```
##     1     1     1     1     1     1     1     1     1     1     2     1     1     1
1
```

```
user_distribution <-ggplot(data=user_count[which(user_count$text.length!=1)
,], aes(x=text.length))

user_distribution +geom_bar(color = "#59C7EB", fill ="#008ECE")+theme_class
ic(base_size=14)+xlab("Tweets per User")+ylab("Amount of Tweets") +ggtitle(
"Distribution of Tweets per User")+theme(axis.text.x = element_text(angle =
90))+theme(text=element_text(family="Times", face="bold", size=16))
```



2. Creating Variables

```
detach("package:dplyr", character.only = TRUE)

library("dplyr", character.only = TRUE)

retweet<-sample %>% select(status_id,retweet_count_log)

colnames(retweet)<-c("text_id","retweet_count")
```

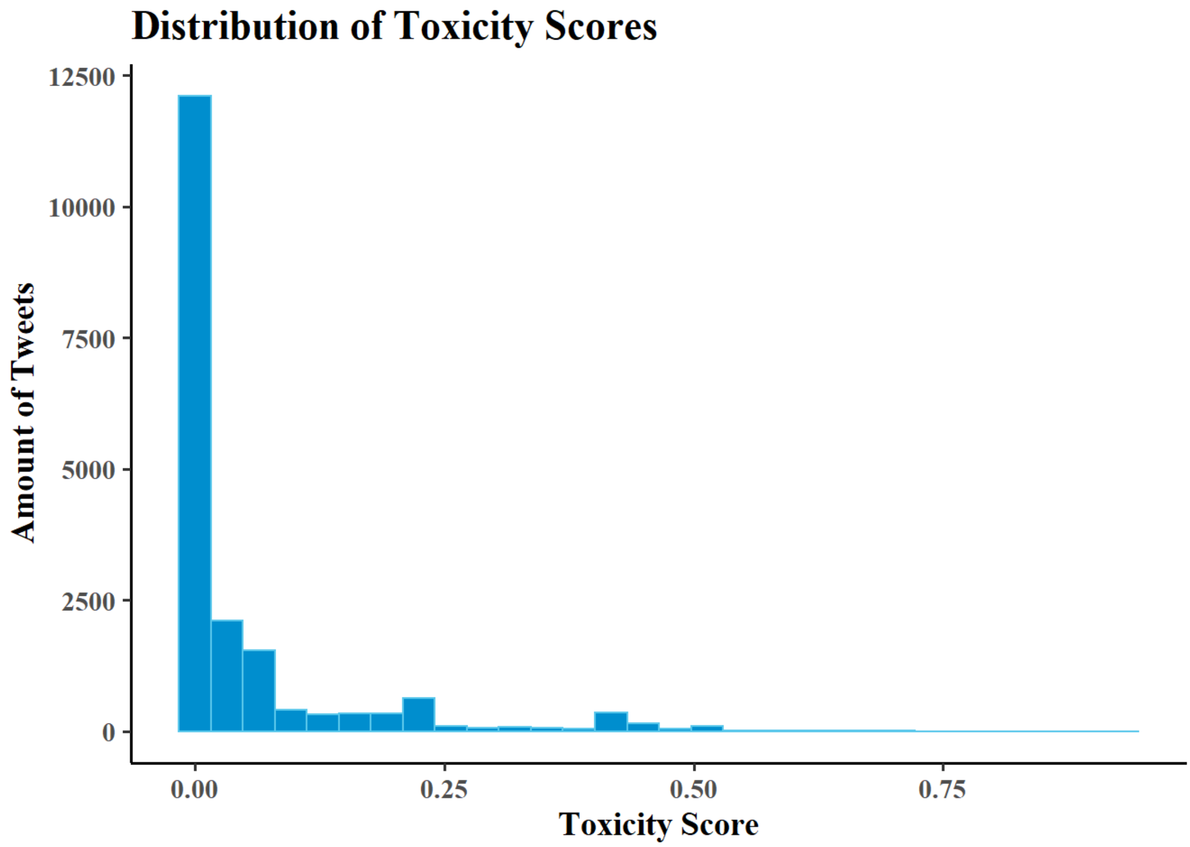
2.0 Control Variables

```
controls <- sample %>% select(status_id,user_id,screen_name,text,hashtags,u
rls_url,followers_count)


controls<-controls%>% mutate(hashtags_dummy = ifelse(hashtags != "NA", 1,0)
)

controls$hashtags_dummy[is.na(controls$hashtags_dummy)] = 0
```

```
controls<-controls%>% mutate(URL_dummy = ifelse(urls_url != "NA", 1,0))

controls$URL_dummy[is.na(controls$URL_dummy)] = 0


controls$followers_count <-log1p(controls$followers_count)

colnames(controls)<-c("text_id", "user_id", "screen_name","text","hashtags"
,"urls_url", "followers_count","hashtags_dummy","URL_dummy")
```

2.1 Determining Toxicity Scores

```
text <- sample %>% select(status_id,text)

colnames(text)<-c("text_id","text")

#DON'T RUN

#connecting to Perspective API


Sys.setenv(perspective_api_key = "XXX")

perspective_api_key = "YYY"


#DON'T RUN

#determine severe toxicity scores

toxicity <-text %>%

prsp_stream(text = text,

                text_id = text_id,

                score_model = "SEVERE_TOXICITY",safe_output=T)


saveRDS(toxicity, "toxicity_scores")

#instead load the Scores

toxicity<-readRDS("toxicity_scores")


toxicity_distribution <-ggplot(data=toxicity, aes(x=SEVERE_TOXICITY))

toxicity_distribution +geom_histogram(color = "#59C7EB", fill ="#008ECE")+t
heme_classic(base_size=14)+xlab("Toxicity Score")+ylab("Amount of Tweets")
+ggtitle("Distribution of Toxicity Scores")+theme(text=element_text(family=
"Times", face="bold", size=14))
```



2.1.1 Binary Toxicity Score

```
#1st quartile value = 0.0030
toxicity$tox_binary[toxicity$SEVERE_TOXICITY >= 0.003]<-1
toxicity$tox_binary[toxicity$SEVERE_TOXICITY < 0.003]<-0
table(toxicity$tox_binary)
```

```
##
##     0     1
##  4663 14163
```

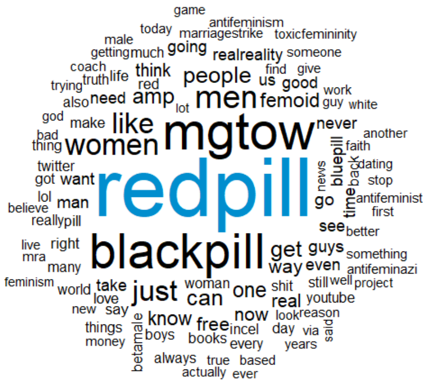
```
#merge toxicity scores with retweet counts
final <- merge(toxicity,retweet, by="text_id")
final<-merge(final,controls,by="text_id")
final<-final %>% select(text_id,user_id,SEVERE_TOXICITY,tox_binary,retweet_
count,URL_dummy,hashtags_dummy,followers_count)
```

2.2 Determining Misogynism Scores

```
#creating a corpus and preparing tokens
corpus <- corpus(text)
text_prepared <-tokens(corpus, remove_punct = T, what = "word1",
                         remove_url =,
                         remove_symbols = T,
                              remove_numbers = T,remove_separators = T,include_docvars = T) %>%
     tokens_tolower %>% tokens_remove(c("#","t.co","https")) %>%
     tokens_remove(stopwords("en"))


#Document-Feature-Matrix
token_dfm <- dfm(text_prepared)


#dictionary of Farrell et al. 2019 (minus the homophobia and racism Category)
misogynisim_dictionary <- dictionary(list(misogynism = c('b00bs', 'becky', 'bint', 'bints', 'bird','birds', 'boob',
  'boobies', 'boobs', 'booobs', 'boooobs', 'booooobs', 'booooooobs', 'chesticles', 'dumb', 'dumbass', 'f4nny', 'failure',
  'fanny', 'fannyflaps', 'female', 'fho', 'fugly', 'funfuck', 'muff', 'pearlnecklace', 'peehole',
  'pissflaps', 'poon', 'poonani', 'poontang', 'pornprincess', 'pua', 'puntang', 'puss', 'pussylips', 'roastie',
  'smv', 'snowflake', 'spermhearder', 'spermherder', 'stacy', 't1tt1e5',
  't1tties', 'tittie', 'titties', 'titty', 'tittyfuck', 'unfuckable', 'va-j-j', 'beta', 'mra', 'normie', 'overthrow', 'prevail', 'vanquish',
  'ass-hat', 'assbag', 'assbite', 'asscock', 'assface', 'asshat', 'asshead', 'asshole', 'assshit', 'asswipe',
  'b!tch', 'b17ch', 'b1tch', 'balls', 'banging', 'bastard', 'beastiality', 'beat', 'beaver',
  'bi+ch', 'biatch', 'bitch', 'bitcher', 'bitchers', 'bitches', 'bitchtits',
  'blockhead', 'blockheads', 'boang', 'bogan', 'bogans', 'bottom-feeder', 'brotherfucker', 'butterhead',
  'butterheads', 'buttface', 'byatch', 'chav', 'chavs',
  'clitface', 'cockbite', 'cockblocker', 'cockhead', 'cockmaster',
  'cockmongler', 'cocknose', 'cocknugget', 'conchuda', 'conchudas', 'coochie', 'coochy', 'crotchrot',
  'cumdumpster', 'cumquat', 'cumqueen', 'cumslut', 'cunt',
  'cuntass', 'cuntface', 'cuntfuck', 'cunthole', 'cuntlick', 'cunts', 'cuntslut', 'demonrats',
  'dickbag', 'dickbrain', 'dickface',
  'dickless', 'dicktickler', 'dickwad', 'dickweed', 'dipshit', 'dipstick',
  'douche', 'douchebag', 'dumbass', 'dumbbitch', 'dumbfuck',
  'entrap', 'ewalt', 'extort', 'fastfuck', 'fatass', 'felcher', 'feltcher', 'fingerfuckers',
  'fistfucker', 'footfucker', 'fucka', 'fuckable', 'fuckass', 'fuckbag', 'fuckboy', 'fuckbrain',
  'fuckbuddy', 'fucker', 'fuckers', 'fuckersucker', 'fuckface', 'fuckfest',
'fuckfreak', 'fuckfriend', 'fuckhead',

  'fuckher', 'fuckina', 'fuckingbitch', 'fuckit', 'fuckknob', 'fuckpig', 'f
ucktard', 'fuckup',

  'fuckwhore', 'fuckyou', 'gangbanger', 'gash', 'gashes', 'greaseball', 'ha
rm', 'hate', 'hayseed', 'hick',

  'hicks', 'hillbilly', 'ho', 'hoar', 'hoare', 'hoe', 'hoer', 'hoes', 'honk
ey', 'honky', 'hoodrat', 'hoodrats',

  'hore', 'hos', 'hurt', 'hussy', 'idiot', 'idiots', 'intimidate', 'jackass
', 'kunt', 'l3i+ch', 'l3itch',

  'lardass', 'libtards', 'limpdick', 'menace', 'milf', 'minge', 'mock',

  'mocks', 'moron', 'mothafuck', 'mothafucka', 'mothafuckas', 'mothafuckaz'
, 'mothafucked', 'mothafucker',

  'mothafuckers', 'mothafucks', 'motherfuck', 'motherfucked', 'motherfucker
', 'motherfuckers', 'motherfuckings',

  'motherfuckka', 'motherfucks', 'muthafecker', 'muthafuckker', 'mutherfuck
er', 'nutsack', 'paleface',

  'palefaces', 'panooch', 'peckerwood', 'pindick', 'pohm', 'pohms', 'pu55i'
, 'pu55y',

  'punish', 'pusse', 'pussi', 'pussie', 'pussies', 'pussy', 'pussys', 'pusy
', 'queerhole', 'redneck', 'rednecks',

  'rentafuck', 'retard', 'retarded', 'russellite', 'russellites', 'scag', '
scags', 'scumbag', 'seppo', 'seppos',

  'sheepfucker', 'sheepfuckers','shitface', 'shithead', 'shitspitter',

  'skag', 'skags', 'skank', 'skanky',  'skullfuck', 'slag',

  'slags', 'slit', 'slits', 'slut', 'slutbag', 'sluts', 'slutt', 'slutting'
, 'slutty',

  'slutwhore', 'smear', 'snatch', 'son-of-a-bitch', 'spermbag', 'suckme', '
suckmytit',

  'tard', 'terrorize', 'threaten', 'thrust', 'titfucker', 'titfuckin', 'tra
ilertrash', 'trisexual', 'turd',

  'tw4t', 'twat', 'twathead', 'twats', 'twatty', 'twatwaffle', 'twobitwhore
', 'twunt', 'twunter', 'wanker',

  'wasp', 'wasps', 'waspy', 'whitey', 'whities', 'whoar', 'whore',

  'whorefucker', 'whores', 'williewanker', 'wuss', 'yankee','amog', 'betabu
xx', 'compel', 'oblige', 'omega', 'overwhelm', 'subjugate', 'suppress',

  'annihilate', 'assail', 'assassinate', 'assault', 'attack', 'bang', 'batt
er', 'blast', 'block',

  'bruise', 'brutalise', 'burn', 'bust', 'butcher', 'choke', 'clobber', 'co
ncuss', 'constrain',

  'crack', 'crush', 'cut', 'decimate', 'demolish', 'destroy','drown', 'ensl
ave', 'er',

  'exterminate', 'flagellate', 'force', 'gag', 'hit', 'jump', 'kick', 'kill
', 'maul',

  'murder', 'obliterate', 'pelt', 'plunk', 'pounce upon', 'pummel', 'punch'
, 'raid', 'ram', 'shake',
  'shoot ', 'shove', 'slam', 'slap', 'slaughter', 'slog', 'smack', 'smash',
'smother', 'stab', 'strangle',
  'strike', 'strong-arm', 'thrash', 'thresh', 'thwack', 'trample', 'trounce
', 'vaporize', 'wallop', 'whip',
  'clitfuck', 'conquer', 'gangbang', 'gangbanged',
  'gangbangs', 'incest', 'infiltrate', 'insest', 'lolita', 'molest', 'moles
tation', 'pound', 'rape', 'sodomise',
  'sodomize', 'spank', 'unclefucker', 'virginbreaker',
  'blackops2cel', 'chad', 'cope', 'cuck', 'currycel', 'fakecel', 'friendles
s', 'fuel', 'gymcel',
  'handholdless', 'heightcel', 'hugless', 'hypergamy', 'incel', 'jbw', 'jfl
', 'kissless', 'kthhfv', 'ldar',
  'looksmatch', 'looksmaxx', 'meeks', 'mogs', 'ricecel', 'rope', 'touchless
', 'truecel', 'tyrone',
  'volcel', 'wagecel', 'wristcel')))


lengths(misogynisim_dictionary)
```

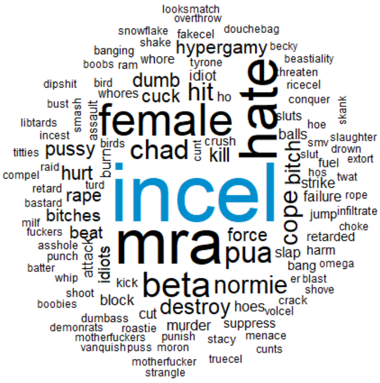
```
## misogynism
##        440
```

```
#applying the dictionary


dict_token <- tokens_lookup(text_prepared, dictionary = misogynisim_diction
ary)
dfm_dict_token<-dfm(dict_token)


dict_words_all<-tokens_keep(text_prepared,pattern=misogynisim_dictionary)
dfm_dict_words_all <- dfm(dict_words_all)


#most frequent words
wordcloud_all <- textplot_wordcloud(token_dfm, min_count = 250, color = "#0
08ECE", random_order = F, ordered_color = T)
```



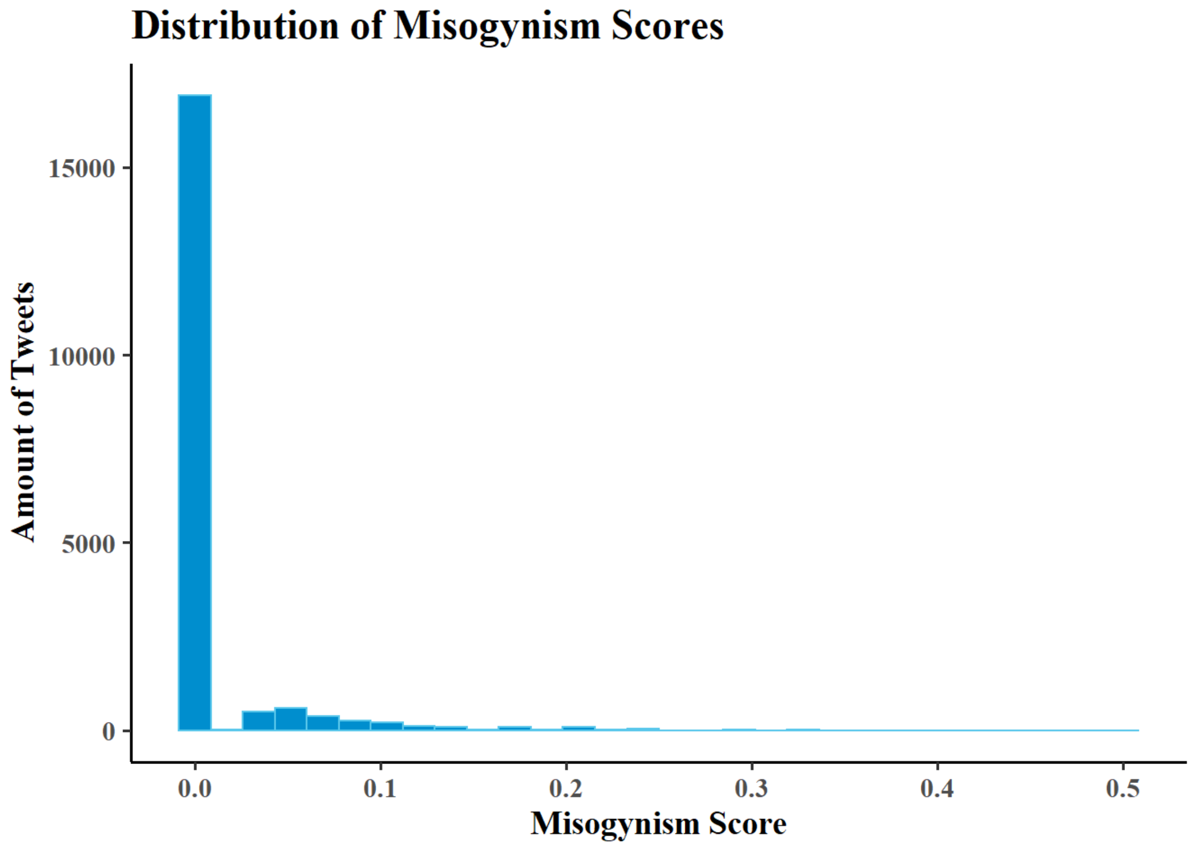
```
topfeatures(token_dfm, n= 20)
##   redpill     mgtow blackpill       men     women      just      like
can
##      6997      3993      3742      2115      1974      1773      1766
1186
##       get    people       amp       one    femoid      know       way
free
##      1158      1153      1100      1013       919       883       819
808
##       now      real     think      want
##       796       776       728       691
#50 most frequent words from the dictionary
topfeatures(dfm_dict_words_all, n= 50)
##     incel       mra      hate    female      beta       pua      chad
cope
##       373       271       224       181       127        99        89
77
##    normie       hit     bitch   destroy      dumb     pussy      hurt
kill
##        74        72        59        59        50        50        48
47
##      cuck hypergamy      rape     force   bitches      beat    idiots
attack
```

```
##         46         46         42         41         36         36         30
28
##      block       burn        cut      balls       hoes       slap      idiot
strike
##         25         25         23         23         23         22         21
21
##     murder    failure         ho   retarded       bang    assault      crush
jump
##         21         20         16         14         13         13         13
12
##       harm      sluts       fuel   suppress       kick      whore     whores
smv
##         12         12         11         10         10         10         10
9
##    conquer     tyrone
##          9          9
wordcloud <- textplot_wordcloud(dfm_dict_words_all, min_count = 2, color =
"#008ECE", random_order = F, ordered_color = T)
```



```
#count of Words by Tweet and Normalisation

tokens_all <- tokens_group(text_prepared, groups = text_id)

ntokens_all <- ntoken(tokens_all)
```

```
misogynist_words <- dfm_group(dfm_dict_token, groups = text_id)
misogynism_scores <- misogynist_words/ntokens_all


#convert dfm to df
misogynism_scores <- convert(misogynism_scores, to = "data.frame")
colnames(misogynism_scores) <- c("text_id", "misogynism_score")


#display Misogynism Scores
misogyny_distribution <-ggplot(data=misogynism_scores, aes(x=misogynism_sco
re))
misogyny_distribution +geom_histogram(color = "#59C7EB", fill ="#008ECE")+t
heme_classic(base_size=14)+xlab("Misogynism Score")+ylab("Amount of Tweets"
) +ggtitle("Distribution of Misogynism Scores")+theme(text=element_text(fam
ily="Times", face="bold", size=14))
```



```
#statistics
mean(misogynism_scores$misogynism_score, na.rm = T)
## [1] 0.01062355
table(misogynism_scores$misogynism_score) #16917 Tweets without any match
##
```

```
##                  0 0.0123456790123457 0.0135135135135135 0.0136986301369
863
##              16917                  1                  1
1
## 0.0144927536231884 0.0147058823529412 0.0149253731343284 0.0153846153846
154
##                  2                  1                  1
1
##           0.015625 0.0161290322580645 0.0163934426229508 0.0166666666666
667
##                  1                  1                  2
1
## 0.0169491525423729 0.0178571428571429 0.0188679245283019 0.0222222222222
222
##                  1                  2                  1
2
## 0.0232558139534884 0.0238095238095238              0.025 0.0256410256410
256
##                  2                  1                  1
1
## 0.0263157894736842  0.027027027027027 0.0285714285714286 0.0294117647058
824
##                  4                  1                  2
6
## 0.0303030303030303            0.03125 0.0317460317460317  0.032258064516
129
##                  9                 11                  1
23
## 0.0333333333333333 0.0344827586206897 0.0357142857142857 0.0363636363636
364
##                 40                 40                 41
1
##  0.037037037037037 0.0384615384615385               0.04 0.0416666666666
667
##                 64                 84                 78
84
## 0.0434782608695652 0.0454545454545455 0.0476190476190476               0
.05
##                 91                 90                 84
86
## 0.0512820512820513 0.0526315789473684 0.0555555555555556 0.0571428571428
571
##                  2                 92                 75
1
## 0.0588235294117647 0.0606060606060606             0.0625 0.0645161290322
581
```

```
##                 71                  3                 78
6

## 0.0666666666666667 0.0689655172413793 0.0714285714285714 0.0740740740740
741

##                 91                  9                 91
20

## 0.0769230769230769               0.08 0.0833333333333333 0.0869565217391
304

##                 81                 19                106
17

## 0.0909090909090909            0.09375 0.0952380952380952
0.1

##                101                  1                 13
101

##  0.103448275862069  0.105263157894737  0.107142857142857  0.111111111111
111

##                  3                 13                  2
80

##  0.115384615384615  0.117647058823529               0.12  0.121212121212
121

##                  1                 12                  8
1

##              0.125  0.130434782608696  0.133333333333333  0.136363636363
636

##                 94                  6                 19
5

##  0.142857142857143  0.148148148148148               0.15  0.153846153846
154

##                 69                  1                  4
9

##  0.157894736842105               0.16  0.166666666666667  0.173913043478
261

##                  3                  1                 93
1

##  0.176470588235294  0.181818181818182             0.1875   0.19047619047
619

##                  3                  8                  2
1

##                0.2  0.208333333333333  0.210526315789474  0.214285714285
714

##                 71                  1                  1
4

##  0.222222222222222  0.230769230769231               0.25  0.272727272727
273

##                  9                  1                 48
1
```

```
##   0.285714285714286                0.3   0.307692307692308   0.333333333333
333
##                   3                  1                   1
15
##                 0.4   0.428571428571429                 0.5
##                   3                  1                   1
```

2.2.1 Binary Misogynism Score

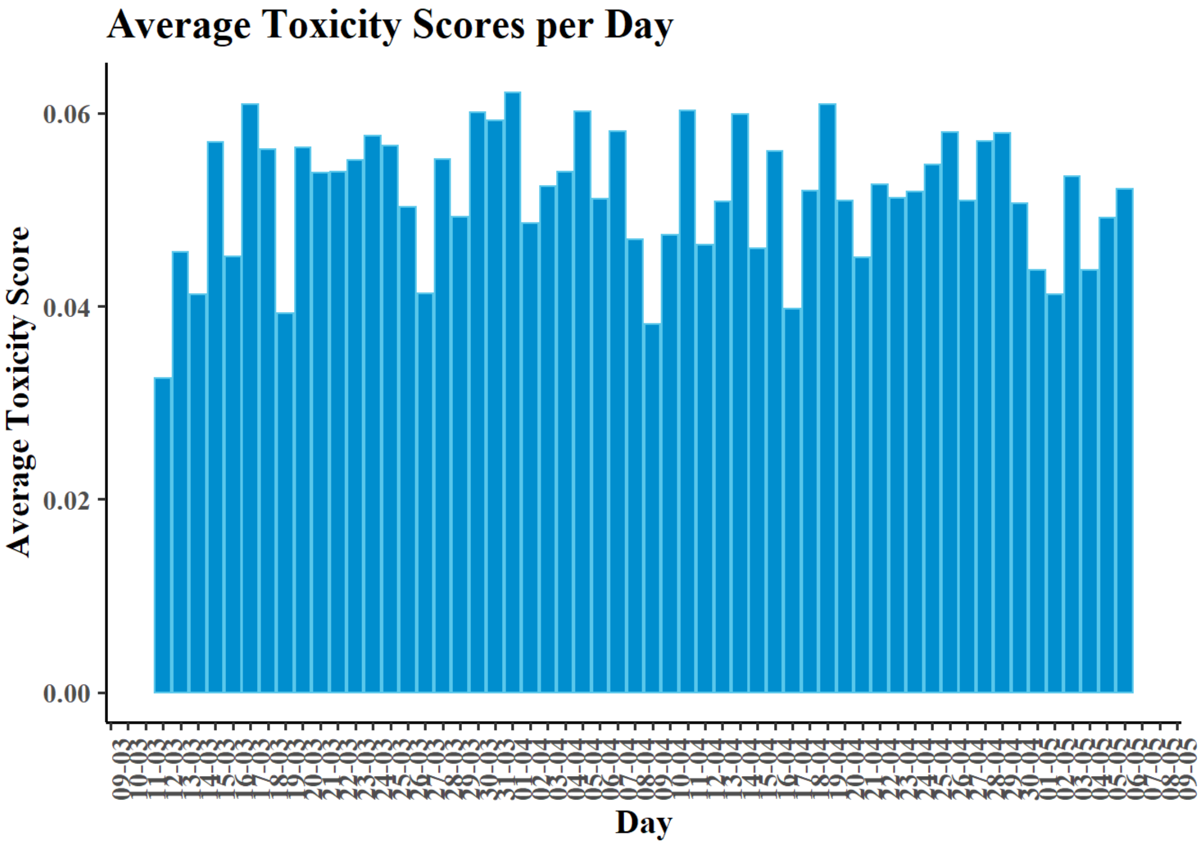
```
misogynism_scores$mis_binary[misogynism_scores$misogynism_score >0]<-1
misogynism_scores$mis_binary[misogynism_scores$misogynism_score ==0]<-0
table(misogynism_scores$mis_binary)
## 
##     0     1 
## 16917  2442
#merge misogynism scores with retweet count
final_2 <- merge(misogynism_scores,retweet, by="text_id")
final_2 <- merge(final_2,controls, by="text_id")
final_2<-final_2 %>% select(text_id,user_id,misogynism_score,mis_binary,ret
weet_count,URL_dummy,hashtags_dummy,followers_count)
```

2.3 Descriptives Toxicity & Misogynism Scores

```
#creating one dataset with both scores
final_all <-merge(final, final_2, by="text_id")
final_tox_mis<-final_all %>% select(text_id,SEVERE_TOXICITY,misogynism_scor
e)


#descriptives toxicity score
final_tox <-toxicity %>% select(SEVERE_TOXICITY)
final_tox<-final_tox_mis %>% select(SEVERE_TOXICITY)


stargazer(final_tox,type="html",out="results/descr_tox.html",covariate.labe
ls=c("Toxicity Score"),title="Descriptive Statistics Toxicity Score")
## 
## <table style="text-align:center"><caption><strong>Descriptive Statistics
Toxicity Score</strong></caption>
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Statistic</td><td>N</td><td>Mean</td><td>St. D
ev.</td><td>Min</td><td>Pctl(25)</td><td>Pctl(75)</td><td>Max</td></tr>
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Toxicity Score</td><td>18,826</td><td>0.052</t
d><td>0.102</td><td>0.00001</td><td>0.003</td><td>0.045</td><td>0.930</td><
/tr>
```

```
## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr></t
able>
```

```
#average toxicity score per day

sample_day <-sample %>% select(status_id, created_at)

colnames(sample_day)<-c("text_id","created_at")

sample_day$day <-format(as.Date(sample_day$created_at,format="%y-%m-%d %h:%
m:%s"),format="%y-%m-%d")

sample_day$day <-as.Date(sample_day$day,format="%y-%m-%d")



tox_day <-merge(final,sample_day,by="text_id") %>% select(text_id, SEVERE_T
OXICITY,day)

tox_av_day <-summaryBy(SEVERE_TOXICITY ~ day, FUN=mean, data=tox_day, na.rm
=TRUE)





average_tox_day <-ggplot(data=tox_av_day, aes(x=day,y=SEVERE_TOXICITY.mean)
)

average_tox_day +geom_col(color = "#59C7EB", fill ="#008ECE")+theme_classic
(base_size=14)+ylab("Average Toxicity Score")+xlab("Day") +ggtitle("Average
Toxicity Scores per Day")+theme(axis.text.x = element_text(angle = 90))+the
me(text=element_text(family="Times", face="bold", size=14))+ scale_x_date(d
ate_breaks = 'day', date_labels = '%d-%m')
```

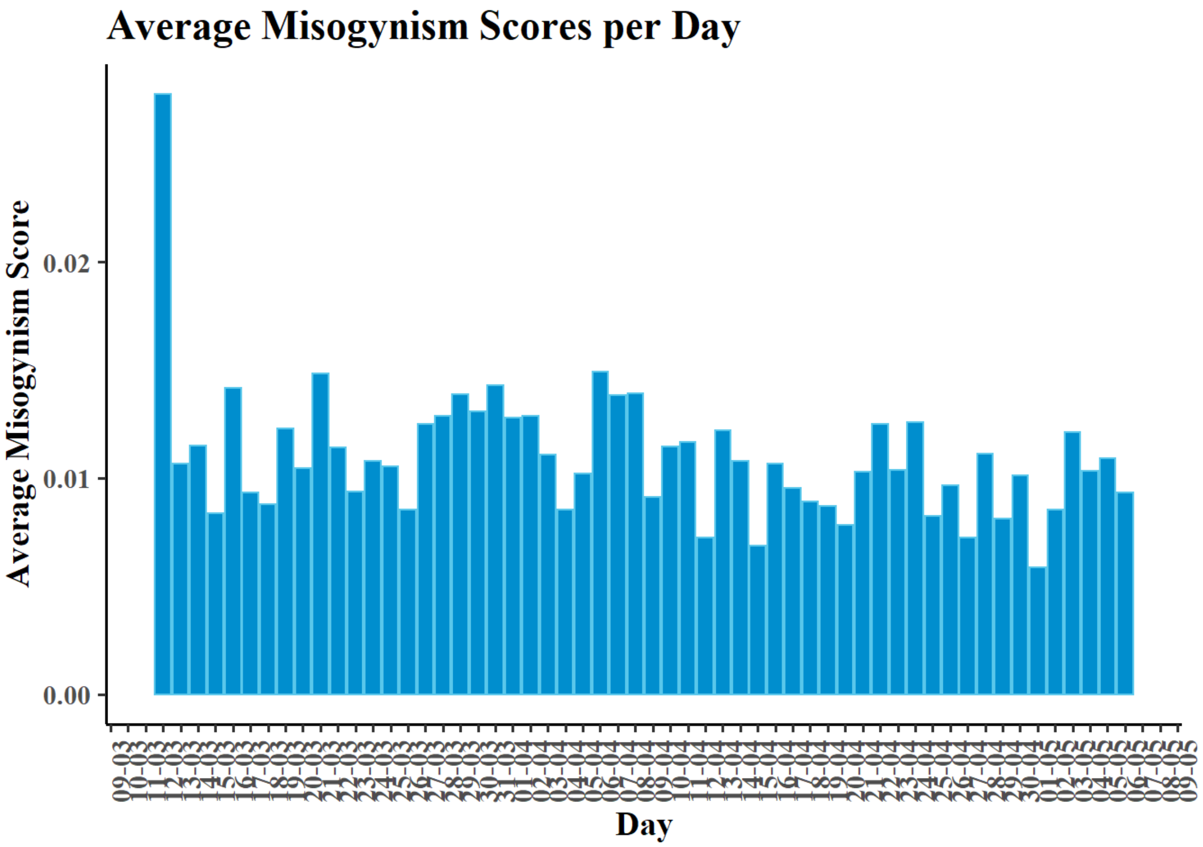


```
#descriptives misogynism score
```

```r
final_mis<-final_tox_mis %>% select(misogynism_score)

stargazer(final_mis,type="html",out="results/descr_mis.html",covariate.labels=c("Misogynism Score"),title="Descriptive Statistics Misogynism Score")
```

```
##

## <table style="text-align:center"><caption><strong>Descriptive Statistics Misogynism Score</strong></caption>

## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Statistic</td><td>N</td><td>Mean</td><td>St. Dev.</td><td>Min</td><td>Pctl(25)</td><td>Pctl(75)</td><td>Max</td></tr>

## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Misogynism Score</td><td>19,359</td><td>0.011</td><td>0.034</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>

## <tr><td colspan="8" style="border-bottom: 1px solid black"></td></tr></table>
```

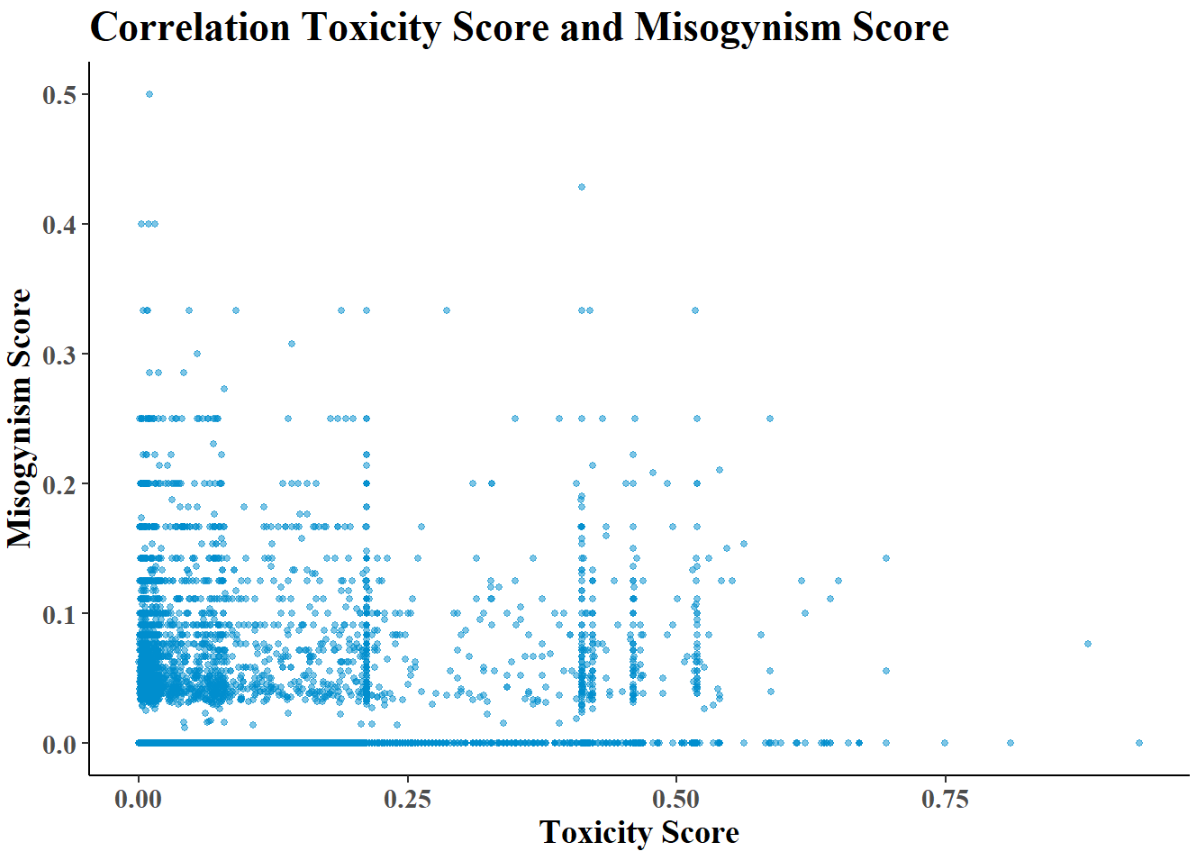
```r
#average misogynism score per day

mis_day <-merge(final_2,sample_day,by="text_id") %>% select(text_id, misogynism_score,day)

mis_av_day <-summaryBy(misogynism_score ~ day, FUN=mean, data=mis_day, na.rm=TRUE)




average_mis_day <-ggplot(data=mis_av_day, aes(x=day,y=misogynism_score.mean))

average_mis_day +geom_col(color = "#59C7EB", fill ="#008ECE")+theme_classic(base_size=14)+ylab("Average Misogynism Score")+xlab("Day") +ggtitle("Average Misogynism Scores per Day")+theme(axis.text.x = element_text(angle = 90))+theme(text=element_text(family="Times", face="bold", size=14))+ scale_x_date(date_breaks = 'day', date_labels = '%d-%m')
```



```
#correlation toxicity & misogynism

cor(final_all$SEVERE_TOXICITY,final_all$misogynism_score, use="complete.obs
")
```

```
## [1] 0.210887
```

```
#scatterplot

tox_mis_scatterplot <-ggplot(data=final_all,mapping=aes(x=SEVERE_TOXICITY,y
= misogynism_score))

tox_mis_scatterplot +geom_point(size=1, alpha=0.5,colour="#008ECE") + theme
_classic() + labs(x = "Toxicity Score", y = "Misogynism Score" )+ ggtitle("
Correlation Toxicity Score and Misogynism Score")+ theme(text=element_text(
family="Times", face = "bold", size=14))
```



3. Analysis

3.1 Analysis H1

```
#correlation retweet and toxicity score

cor(final$retweet_count,final$SEVERE_TOXICITY, use="complete.obs")
```

```
## [1] -0.007410595
```

```
#model construction

model_1<-lm(retweet_count ~ SEVERE_TOXICITY,data=final , na.action = na.omi
t)

model_2<-lm(retweet_count ~tox_binary, data=final,na.action = na.omit)

model_3<-plm(retweet_count~SEVERE_TOXICITY +URL_dummy +hashtags_dummy +foll
owers_count, index=c("user_id"), data=final, model="within",effect = "indiv
idual",na.action = na.omit)

model_4<-plm(retweet_count~tox_binary +URL_dummy +hashtags_dummy +followers
_count, index=c("user_id"), data=final, model="within",effect = "individual
",na.action = na.omit)



#output models

summary(model_1)
```

```
##

## Call:
```

```
## lm(formula = retweet_count ~ SEVERE_TOXICITY, data = final, na.action =
na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2161 -0.2159 -0.2155 -0.2080  7.0455
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.216079   0.004918  43.936   <2e-16 ***
## SEVERE_TOXICITY -0.043558   0.042840  -1.017    0.309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.602 on 18824 degrees of freedom
##   (533 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  5.492e-05,  Adjusted R-squared:  1.796e-06
## F-statistic: 1.034 on 1 and 18824 DF,  p-value: 0.3093
```

```
summary(model_2)
```

```
##
## Call:
## lm(formula = retweet_count ~ tox_binary, data = final, na.action = na.om
it)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2463 -0.2031 -0.2031 -0.2031  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.246347   0.008812  27.957  < 2e-16 ***
## tox_binary  -0.043236   0.010159  -4.256 2.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6017 on 18824 degrees of freedom
##   (533 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.0009613,  Adjusted R-squared:  0.0009082
## F-statistic: 18.11 on 1 and 18824 DF,  p-value: 2.092e-05
```

```
summary(model_3)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = retweet_count ~ SEVERE_TOXICITY + URL_dummy + hashtags_dum
my +
##     followers_count, data = final, na.action = na.omit, effect = "indivi
dual",
##     model = "within", index = c("user_id"))
##
## Unbalanced Panel: n = 10498, T = 1-273, N = 18826
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -3.212055 -0.013652  0.000000  0.000000  4.611029
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## SEVERE_TOXICITY 0.1824013  0.0591414  3.0842  0.002048 **
## URL_dummy       0.0086018  0.0185842  0.4629  0.643479
## hashtags_dummy  0.2090068  0.0259259  8.0617 8.568e-16 ***
## followers_count 0.1769246  0.0263477  6.7150 2.004e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1801.7
## Residual Sum of Squares: 1775.3
## R-Squared:      0.01463
## Adj. R-Squared: -1.2284
## F-statistic: 30.8969 on 4 and 8324 DF, p-value: < 2.22e-16
```

```
summary(model_4)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = retweet_count ~ tox_binary + URL_dummy + hashtags_dummy +
##     followers_count, data = final, na.action = na.omit, effect = "indivi
dual",
##     model = "within", index = c("user_id"))
```

```
##
## Unbalanced Panel: n = 10498, T = 1-273, N = 18826
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -3.206384 -0.012376  0.000000  0.000000  4.618462
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## tox_binary      0.040838   0.014199  2.8762  0.004035 **
## URL_dummy       0.013077   0.018644  0.7014  0.483067
## hashtags_dummy  0.210723   0.025930  8.1267 5.047e-16 ***
## followers_count 0.174819   0.026363  6.6312 3.538e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1801.7
## Residual Sum of Squares: 1775.6
## R-Squared:      0.014483
## Adj. R-Squared: -1.2288
## F-statistic: 30.5827 on 4 and 8324 DF, p-value: < 2.22e-16
```

```
#creating a formatted table
stargazer(model_1,model_2,model_3, model_4,type= "html",out="results/tox_results.html",title="Regression Results H1:",dep.var.labels = "Retweet Count",covariate.labels = c("Toxicity Score","Toxicity Binary","Presence of URLs","Presence of Hashtags","Number of Followers"))
```

```
##
## <table style="text-align:center"><caption><strong>Regression Results H1:</strong></caption>
## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"></td><td colspan="4"><em>Dependent variable:</em></td></tr>
## <tr><td></td><td colspan="4" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td colspan="4">Retweet Count</td></tr>
## <tr><td style="text-align:left"></td><td colspan="2"><em>OLS</em></td><td colspan="2"><em>panel</em></td></tr>
## <tr><td style="text-align:left"></td><td colspan="2"><em></em></td><td colspan="2"><em>linear</em></td></tr>
## <tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td><td>(4)</td></tr>
## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Toxicity Score</td><td>-0.044</td><td></td><td
>0.182<sup>***</sup></td><td></td></tr>

## <tr><td style="text-align:left"></td><td>(0.043)</td><td></td><td>(0.059
)</td><td></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Toxicity Binary</td><td></td><td>-0.043<
sup>***</sup></td><td></td><td>0.041<sup>***</sup></td></tr>

## <tr><td style="text-align:left"></td><td></td><td>(0.010)</td><td></td><
td>(0.014)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Presence of URLs</td><td></td><td></td><
td>0.009</td><td>0.013</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.019)</td><
td>(0.019)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Presence of Hashtags</td><td></td><td></
td><td>0.209<sup>***</sup></td><td>0.211<sup>***</sup></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.026)</td><
td>(0.026)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Number of Followers</td><td></td><td></t
d><td>0.177<sup>***</sup></td><td>0.175<sup>***</sup></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.026)</td><
td>(0.026)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Constant</td><td>0.216<sup>***</sup></td
><td>0.246<sup>***</sup></td><td></td><td></td></tr>

## <tr><td style="text-align:left"></td><td>(0.005)</td><td>(0.009)</td><td
></td><td></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Observations</td><td>18,826</td><td>18,826</td
><td>18,826</td><td>18,826</td></tr>

## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.0001</td><td>0.0
01</td><td>0.015</td><td>0.014</td></tr>

## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.00000</
td><td>0.001</td><td>-1.228</td><td>-1.229</td></tr>

## <tr><td style="text-align:left">Residual Std. Error (df = 18824)</td><td
>0.602</td><td>0.602</td><td></td><td></td></tr>
```

```
## <tr><td style="text-align:left">F Statistic</td><td>1.034 (df = 1; 18824
)</td><td>18.113<sup>***</sup> (df = 1; 18824)</td><td>30.897<sup>***</sup>
(df = 4; 8324)</td><td>30.583<sup>***</sup> (df = 4; 8324)</td></tr>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left"><em>Note:</em></td><td colspan="4" style="text
-align:right"><sup>*</sup>p<0.1; <sup>**</sup>p<0.05; <sup>***</sup>p<0.01<
/td></tr>

## </table>
```

```
#creating standardised coefficients

zmodel_3<-lm.beta(model_3)

stargazer(zmodel_3, type="html",out="results/ztox_3.html",covariate.labels
= c("Toxicity Score","Presence of URLs","Presence of Hashtags","Number of F
ollowers"),title="Standardised Coefficients Model 3")
```

```
##

## <table style="text-align:center"><caption><strong>Standardised Coefficie
nts Model 3</strong></caption>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Toxicity Score</td><td>Presence of URLs</td><t
d>Presence of Hashtags</td><td>Number of Followers</td></tr>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">0.001</td><td>0.149</td><td>0.127</td><td>0.03
2</td></tr>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr></t
able>
```

```
zmodel_4<-lm.beta(model_4)

stargazer(zmodel_4, type="html",out="results/ztox_4.html",covariate.labels
= c("Toxicity Binary","Presence of URLs","Presence of Hashtags","Number of
Followers"),title="Standardised Coefficients Model 4")
```

```
##

## <table style="text-align:center"><caption><strong>Standardised Coefficie
nts Model 4</strong></caption>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Toxicity Binary</td><td>Presence of URLs</td><
td>Presence of Hashtags</td><td>Number of Followers</td></tr>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">0.009</td><td>0.150</td><td>0.125</td><td>0.04
9</td></tr>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr></t
able>
```

3.1.1 Robustness Checks

3.1.1.1 Homogeneity of Variance

```
#Analytic: Breusch-Pagan Test



#model 1 and 2 are nor assessed as the estimates are not significant or in
the wrong direction
```

```
#Model 3
bptest(model_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 901.85, df = 4, p-value < 2.2e-16
```

```
#H0: Homogeneity
#H1: Heterogeneity
#p < 0.05 = Heterogeneity


#determining robust standard errors
model_3_robust <-coeftest(model_3,vcov=vcovHC(model_3))


stargazer(model_3_robust,type="html",out="results/robust_3.html",title="Robust Standard Errors Model 3:",dep.var.labels = "Retweet Count",covariate.labels = c("Toxicity Score","Presence of URLs","Presence of Hashtags","Number of Followers","Age of the Account"))
```

```
##
## <table style="text-align:center"><caption><strong>Robust Standard Errors Model 3:</strong></caption>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left"></td><td><em>Dependent variable:</em></td></tr>
## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td></tr>
## <tr><td style="text-align:left"></td><td>Retweet Count</td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Toxicity Score</td><td>0.182<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.064)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Presence of URLs</td><td>0.009</td></tr>
## <tr><td style="text-align:left"></td><td>(0.050)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Presence of Hashtags</td><td>0.209<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.061)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Number of Followers</td><td>0.177<sup>***</sup></td></tr>
## <tr><td style="text-align:left"></td><td>(0.050)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td s
tyle="text-align:left"><em>Note:</em></td><td style="text-align:right"><sup
>*</sup>p<0.1; <sup>**</sup>p<0.05; <sup>***</sup>p<0.01</td></tr>

## </table>
```

```
#Model 4

bptest(model_4)
```

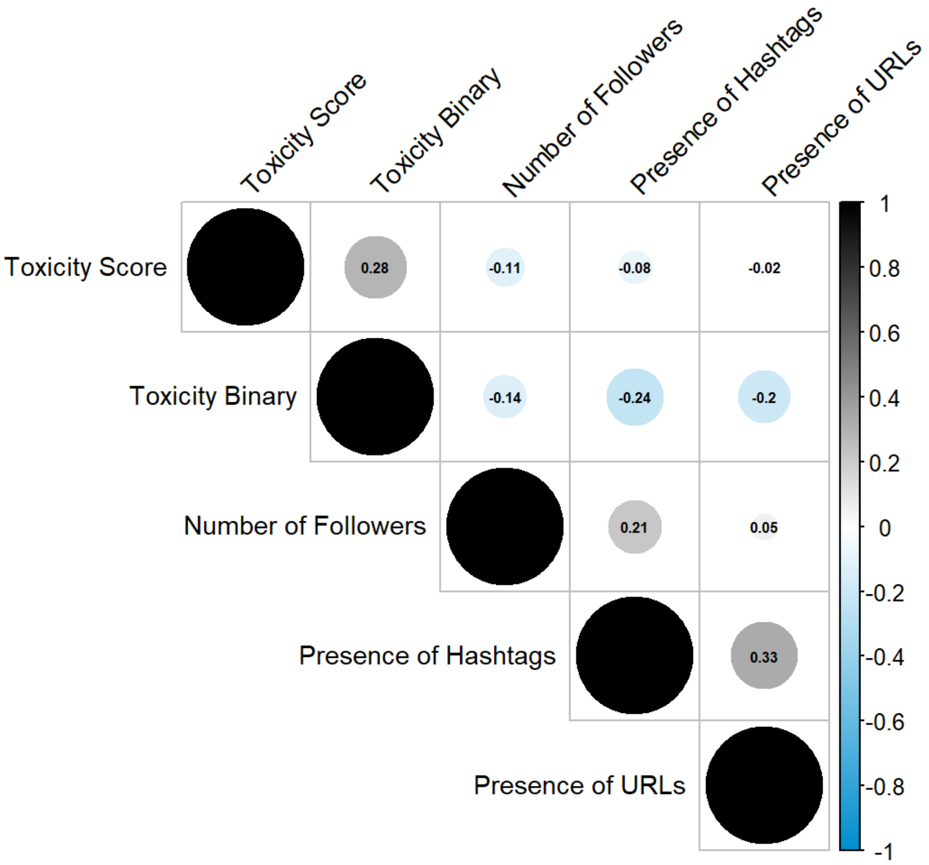
```
##

##  studentized Breusch-Pagan test

##

## data:  model_4

## BP = 902.08, df = 4, p-value < 2.2e-16
```

```
#H0: Homogeneity

#H1: Heterogeneity

#p < 0.05 = Heterogeneity


#determining robust standard errors

model_4_robust <-coeftest(model_4,vcov=vcovHC(model_4))


stargazer(model_4_robust,type="html",out="results/robust_4.html",title="Rob
ust Standard Errors Model 4:",dep.var.labels = "Retweet Count",covariate.la
bels = c("Toxicity Binary","Presence of URLs","Presence of Hashtags","Numbe
r of Followers","Age of the Account"))
```

```
##

## <table style="text-align:center"><caption><strong>Robust Standard Errors
Model 4:</strong></caption>

## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left"></td><td><em>Dependent variable:</em></td></tr
>

## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td
></tr>

## <tr><td style="text-align:left"></td><td>Retweet Count</td></tr>

## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Toxicity Binary</td><td>0.041<sup>*</sup></td>
</tr>

## <tr><td style="text-align:left"></td><td>(0.025)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td style="text-align:left">Presence of URLs</td><td>0.013</td></tr>

## <tr><td style="text-align:left"></td><td>(0.047)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Presence of Hashtags</td><td>0.211<sup>*
**</sup></td></tr>

## <tr><td style="text-align:left"></td><td>(0.061)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td style="text-align:left">Number of Followers</td><td>0.175<sup>**
*</sup></td></tr>

## <tr><td style="text-align:left"></td><td>(0.049)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td s
tyle="text-align:left"><em>Note:</em></td><td style="text-align:right"><sup
>*</sup>p<0.1; <sup>**</sup>p<0.05; <sup>***</sup>p<0.01</td></tr>

## </table>
```

3.1.1.2 Excluding perfect Mulitcollinearity

```
cor_final <-final %>% select(SEVERE_TOXICITY,tox_binary,URL_dummy,hashtags_
dummy,followers_count)

cor_final <-round(cor_final,3)

colnames(cor_final)<-c("Toxicity Score","Toxicity Binary", "Presence of Has
htags","Presence of URLs","Number of Followers")


cor_h1<-cor(cor_final,use="pairwise.complete.obs")


corrplot(cor_h1,type="upper", tl.col = "black", tl.srt = 45,order="hclust",
col=colorRampPalette(c("#008ECE","white","black"))(200),addCoef.col = 1,num
ber.cex = 0.5,tl.cex=0.9)
```



3.1.1.3 Fixed Effects vs. Random Effects

```
#Hausmantest


#Model 3

FEmodel<-model_3<-plm(retweet_count~SEVERE_TOXICITY +URL_dummy +hashtags_du
mmy +followers_count, index=c("user_id"), data=final, model="within",effect
="individual")

REmodel<-model_3<-plm(retweet_count~SEVERE_TOXICITY +URL_dummy +hashtags_du
mmy +followers_count, index=c("user_id"), data=final, model="random",effect
="individual")


#H0: RE model is consistent

#H1: FE model is consistent

phtest(FEmodel,REmodel)
```

```
##

##  Hausman Test

##

## data:  retweet_count ~ SEVERE_TOXICITY + URL_dummy + hashtags_dummy +  .
..

## chisq = 22.485, df = 4, p-value = 0.0001604

## alternative hypothesis: one model is inconsistent
```

```
#p-value = 4.435e-05 and thus below 0.05, thus FE is chosen
```

```
#Model 4

FEmodel<-model_4<-plm(retweet_count~tox_binary +URL_dummy +hashtags_dummy +
followers_count, index=c("user_id"), data=final, model="within",effect="ind
ividual")

REmodel<-model_4<-plm(retweet_count~tox_binary +URL_dummy +hashtags_dummy +
followers_count, index=c("user_id"), data=final, model="random",effect="ind
ividual")


#H0: RE model is consistent

#H1: FE model is consistent

phtest(FEmodel,REmodel)

##

##  Hausman Test

##

## data:  retweet_count ~ tox_binary + URL_dummy + hashtags_dummy + followe
rs_count

## chisq = 24.78, df = 4, p-value = 5.571e-05

## alternative hypothesis: one model is inconsistent

#p-value = 1.988e-05 and thus below 0.05, thus FE is chosen
```

3.1.1.4 Independence of Errors

```
#Durbin Watson Test

#D-W values must be as close as possible to 2 but at least in between 1.5 a
nd 2.5

durbinWatsonTest(model_1)

##  lag Autocorrelation D-W Statistic p-value

##    1     0.006147996      1.987691    0.39

##  Alternative hypothesis: rho != 0

durbinWatsonTest(model_2)

##  lag Autocorrelation D-W Statistic p-value

##    1     0.006338103      1.987312   0.368

##  Alternative hypothesis: rho != 0

pdwtest(model_3)

##

##  Durbin-Watson test for serial correlation in panel models

##

## data:  retweet_count ~ SEVERE_TOXICITY + URL_dummy + hashtags_dummy +
followers_count

## DW = 1.9318, p-value = 1.402e-06

## alternative hypothesis: serial correlation in idiosyncratic errors
```

```
pdwtest(model_4)
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  retweet_count ~ tox_binary + URL_dummy + hashtags_dummy + followe
rs_count
## DW = 1.9319, p-value = 1.455e-06
## alternative hypothesis: serial correlation in idiosyncratic errors
```

3.2 Analysis H2

```
#correlation retweet and misogynism score
cor(final_2$misogynism_score,final$retweet_count, use="complete.obs")
## [1] -0.009231137
model_5<-lm(retweet_count ~ misogynism_score,data=final_2,na.action = na.om
it )
model_6<-lm(retweet_count ~ mis_binary,data=final_2,na.action = na.omit )
model_7 <-plm(retweet_count ~misogynism_score + URL_dummy + hashtags_dummy
+ followers_count,index="user_id",model="within", data=final_2, na.action =
na.omit)
model_8<-plm(retweet_count ~mis_binary + URL_dummy + hashtags_dummy + follo
wers_count,index="user_id",model="within", data=final_2, na.action = na.omi
t)


#output models
summary(model_5)
##
## Call:
## lm(formula = retweet_count ~ misogynism_score, data = final_2,
##     na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2153 -0.2153 -0.2153 -0.2086  7.0459
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.215306   0.004529  47.536   <2e-16 ***
## misogynism_score -0.161571   0.125797  -1.284    0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.6021 on 19357 degrees of freedom
## Multiple R-squared:  8.521e-05,  Adjusted R-squared:  3.356e-05
## F-statistic:  1.65 on 1 and 19357 DF,  p-value: 0.199
```

```
summary(model_6)
```

```
##
## Call:
## lm(formula = retweet_count ~ mis_binary, data = final_2, na.action = na.
omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2246 -0.2120 -0.2120 -0.2120  7.0492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.21200    0.00463  45.793   <2e-16 ***
## mis_binary   0.01259    0.01303   0.966    0.334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6022 on 19357 degrees of freedom
## Multiple R-squared:  4.818e-05,  Adjusted R-squared:  -3.479e-06
## F-statistic: 0.9327 on 1 and 19357 DF,  p-value: 0.3342
```

```
summary(model_7)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = retweet_count ~ misogynism_score + URL_dummy +
##     hashtags_dummy + followers_count, data = final_2, na.action = na.omi
t,
##     model = "within", index = "user_id")
##
## Unbalanced Panel: n = 10789, T = 1-273, N = 19359
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -3.2063755 -0.0082797  0.0000000  0.0000000  4.6183738
```

```
##
## Coefficients:
##                  Estimate Std. Error t-value  Pr(>|t|)
## misogynism_score 0.1656926  0.1770803  0.9357    0.3495
## URL_dummy        0.0052318  0.0183366  0.2853    0.7754
## hashtags_dummy   0.2008863  0.0255847  7.8518 4.595e-15 ***
## followers_count  0.1790149  0.0260459  6.8731 6.723e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1836.5
## Residual Sum of Squares: 1812.6
## R-Squared:      0.013025
## Adj. R-Squared: -1.2304
## F-statistic: 28.2611 on 4 and 8566 DF, p-value: < 2.22e-16
```

```
summary(model_8)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = retweet_count ~ mis_binary + URL_dummy + hashtags_dummy +
##     followers_count, data = final_2, na.action = na.omit, model = "withi
n",
##     index = "user_id")
##
## Unbalanced Panel: n = 10789, T = 1-273, N = 19359
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -3.206374 -0.011222  0.000000  0.000000  4.618362
##
## Coefficients:
##                 Estimate Std. Error t-value  Pr(>|t|)
## mis_binary      0.0259979  0.0170282  1.5268    0.1269
## URL_dummy       0.0054139  0.0183357  0.2953    0.7678
## hashtags_dummy  0.2008029  0.0255817  7.8495 4.681e-15 ***
## followers_count 0.1795786  0.0260466  6.8945 5.788e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##

## Total Sum of Squares:    1836.5

## Residual Sum of Squares: 1812.3

## R-Squared:      0.013193

## Adj. R-Squared: -1.2301

## F-statistic: 28.6297 on 4 and 8566 DF, p-value: < 2.22e-16
```

```
#creating a formatted table

stargazer(model_5,model_6,model_7, model_8,type= "html",out="results/mis_re
sults.html",title="Regression Results H2:",dep.var.labels = "Retweet Count"
,covariate.labels = c("Misogynism Score","Misogynism Binary","Presence of U
RLs","Presence of Hashtags","Number of Followers"))
```

```
##

## <table style="text-align:center"><caption><strong>Regression Results H2:
</strong></caption>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left"></td><td colspan="4"><em>Dependent variable:</
em></td></tr>

## <tr><td></td><td colspan="4" style="border-bottom: 1px solid black"></td
></tr>

## <tr><td style="text-align:left"></td><td colspan="4">Retweet Count</td><
/tr>

## <tr><td style="text-align:left"></td><td colspan="2"><em>OLS</em></td><t
d colspan="2"><em>panel</em></td></tr>

## <tr><td style="text-align:left"></td><td colspan="2"><em></em></td><td c
olspan="2"><em>linear</em></td></tr>

## <tr><td style="text-align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td
><td>(4)</td></tr>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Misogynism Score</td><td>-0.162</td><td></td><
td>0.166</td><td></td></tr>

## <tr><td style="text-align:left"></td><td>(0.126)</td><td></td><td>(0.177
)</td><td></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Misogynism Binary</td><td></td><td>0.013
</td><td></td><td>0.026</td></tr>

## <tr><td style="text-align:left"></td><td></td><td>(0.013)</td><td></td><
td>(0.017)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Presence of URLs</td><td></td><td></td><
td>0.005</td><td>0.005</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.018)</td><
td>(0.018)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>
```

```
## <tr><td style="text-align:left">Presence of Hashtags</td><td></td><td></
td><td>0.201<sup>***</sup></td><td>0.201<sup>***</sup></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.026)</td><
td>(0.026)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Number of Followers</td><td></td><td></t
d><td>0.179<sup>***</sup></td><td>0.180<sup>***</sup></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td>(0.026)</td><
td>(0.026)</td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td style="text-align:left">Constant</td><td>0.215<sup>***</sup></td
><td>0.212<sup>***</sup></td><td></td><td></td></tr>

## <tr><td style="text-align:left"></td><td>(0.005)</td><td>(0.005)</td><td
></td><td></td></tr>

## <tr><td style="text-align:left"></td><td></td><td></td><td></td><td></td
></tr>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Observations</td><td>19,359</td><td>19,359</td
><td>19,359</td><td>19,359</td></tr>

## <tr><td style="text-align:left">R<sup>2</sup></td><td>0.0001</td><td>0.0
0005</td><td>0.013</td><td>0.013</td></tr>

## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>0.00003</
td><td>-0.00000</td><td>-1.230</td><td>-1.230</td></tr>

## <tr><td style="text-align:left">Residual Std. Error (df = 19357)</td><td
>0.602</td><td>0.602</td><td></td><td></td></tr>

## <tr><td style="text-align:left">F Statistic</td><td>1.650 (df = 1; 19357
)</td><td>0.933 (df = 1; 19357)</td><td>28.261<sup>***</sup> (df = 4; 8566)
</td><td>28.630<sup>***</sup> (df = 4; 8566)</td></tr>

## <tr><td colspan="5" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left"><em>Note:</em></td><td colspan="4" style="text
-align:right"><sup>*</sup>p<0.1; <sup>**</sup>p<0.05; <sup>***</sup>p<0.01<
/td></tr>

## </table>
```

```
##standardised coefficients for model 8

zmodel_8<-lm.beta(model_8)

stargazer(zmodel_8, type="html",out="results/zmis_8.html",covariate.labels
= c("Misogynism Binary","Presence of URLs","Presence of Hashtags","Number o
f Followers"),title="Standardised Coefficients Model 8")
```

```
##

## <table style="text-align:center"><caption><strong>Standardised Coefficie
nts Model 8</strong></caption>

## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Misogynism Binary</td><td>Presence of URLs</td
><td>Presence of Hashtags</td><td>Number of Followers</td></tr>
```

```
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">0.003</td><td>0.142</td><td>0.128</td><td>0.02
0</td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid black"></td></tr></t
able>
```

3.2.1 Robustness Checks

3.2.1.1 Homogeneity of Variance

```
#Analytic: Breusch-Pagan Test

#Only model 8 is assesed
bptest(model_8)
```

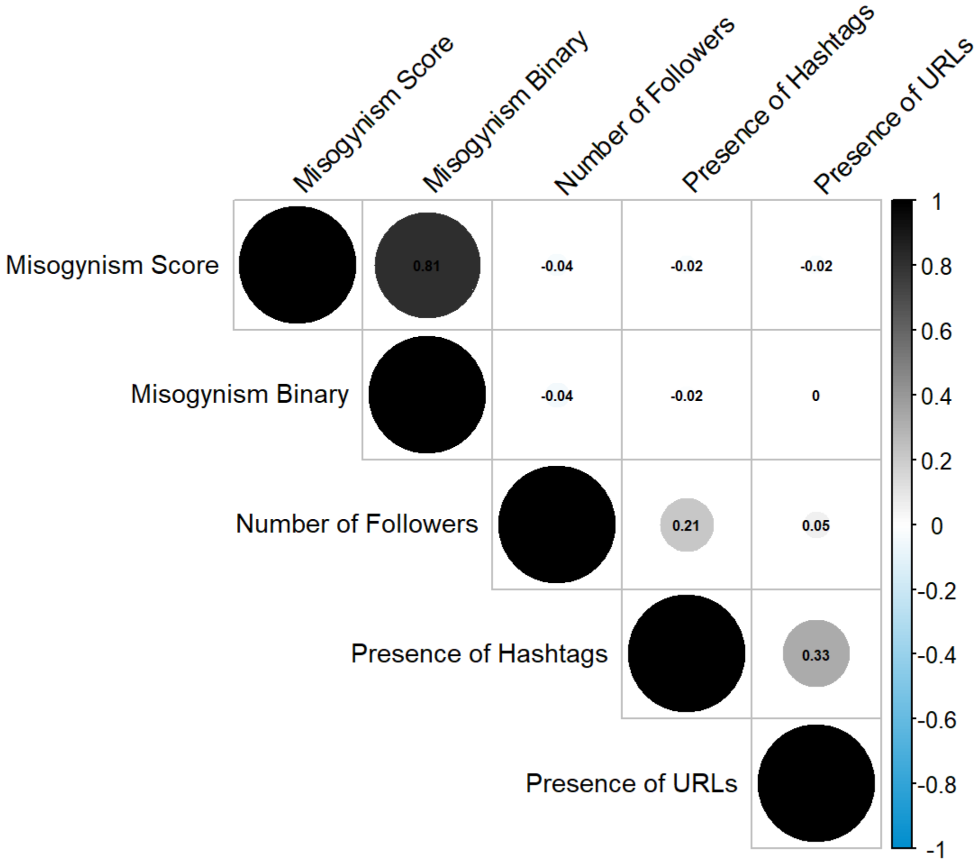
```
##
##  studentized Breusch-Pagan test
##
## data:  model_8
## BP = 936.44, df = 4, p-value < 2.2e-16
```

```
#H0: Homogeneity
#H1: Heterogeneity
#p-value = < 2.2e-16 = Heterogeneity

#determining robust standard errors
model_8_robust <-coeftest(model_8,vcov=vcovHC(model_8))

stargazer(model_8_robust,type="html",out="results/robust_8.html",title="Rob
ust Standard Errors Model 8:",dep.var.labels = "Retweet Count",covariate.la
bels = c("Misogynism Binary","Presence of URLs","Presence of Hashtags","Num
ber of Followers"))
```

```
##
## <table style="text-align:center"><caption><strong>Robust Standard Errors
Model 8:</strong></caption>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left"></td><td><em>Dependent variable:</em></td></tr
>
## <tr><td></td><td colspan="1" style="border-bottom: 1px solid black"></td
></tr>
## <tr><td style="text-align:left"></td><td>Retweet Count</td></tr>
## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td style="text-align:left">Misogynism Binary</td><td>0.026</td></tr>
## <tr><td style="text-align:left"></td><td>(0.019)</td></tr>
## <tr><td style="text-align:left"></td><td></td></tr>
## <tr><td style="text-align:left">Presence of URLs</td><td>0.005</td></tr>
## <tr><td style="text-align:left"></td><td>(0.048)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td style="text-align:left">Presence of Hashtags</td><td>0.201<sup>*
**</sup></td></tr>

## <tr><td style="text-align:left"></td><td>(0.060)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td style="text-align:left">Number of Followers</td><td>0.180<sup>**
*</sup></td></tr>

## <tr><td style="text-align:left"></td><td>(0.049)</td></tr>

## <tr><td style="text-align:left"></td><td></td></tr>

## <tr><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr
><td colspan="2" style="border-bottom: 1px solid black"></td></tr><tr><td s
tyle="text-align:left"><em>Note:</em></td><td style="text-align:right"><sup
>*</sup>p<0.1; <sup>**</sup>p<0.05; <sup>***</sup>p<0.01</td></tr>

## </table>
```

3.2.1.2 Excluding perfect Mulitcollinearity

```
cor_final_2 <-final_2 %>% select(misogynism_score,mis_binary,URL_dummy,hash
tags_dummy,followers_count)

cor_final_2 <-round(cor_final_2,3)

colnames(cor_final_2)<-c("Misogynism Score","Misogynism Binary", "Presence
of Hashtags","Presence of URLs","Number of Followers")


cor_h2<-cor(cor_final_2,use="pairwise.complete.obs")


corrplot(cor_h2,type="upper", tl.col = "black", tl.srt = 45,order="hclust",
col=colorRampPalette(c("#008ECE","white","black"))(200),addCoef.col = 1,num
ber.cex = 0.5,tl.cex=0.9)
```



3.2.1.3 Fixed Effects vs. Random Effects

```
#Hausmantest


#Model 7

FEmodel<-model_7<-plm(retweet_count~misogynism_score +URL_dummy +hashtags_d
ummy +followers_count, index=c("user_id"), data=final_2, model="within",eff
ect="individual")

REmodel<-model_7<-plm(retweet_count~misogynism_score +URL_dummy +hashtags_d
ummy +followers_count, index=c("user_id"), data=final_2, model="random",eff
ect="individual")


#H0: RE model is consistent

#H1: FE model is consistent

phtest(FEmodel,REmodel)

##

##  Hausman Test

##

## data:  retweet_count ~ misogynism_score + URL_dummy + hashtags_dummy +
...

## chisq = 22.333, df = 4, p-value = 0.0001721

## alternative hypothesis: one model is inconsistent

#p-value = 0.0001571 and thus below 0.05, thus FE is chosen
```

```
#Model 8

FEmodel<-model_8<-plm(retweet_count~mis_binary +URL_dummy +hashtags_dummy +
followers_count, index=c("user_id"), data=final_2, model="within",effect="i
ndividual")

REmodel<-model_8<-plm(retweet_count~mis_binary +URL_dummy +hashtags_dummy +
followers_count, index=c("user_id"), data=final_2, model="random",effect="i
ndividual")



#H0: RE model is consistent

#H1: FE model is consistent

phtest(FEmodel,REmodel)

##

##  Hausman Test

##

## data:  retweet_count ~ mis_binary + URL_dummy + hashtags_dummy + followe
rs_count

## chisq = 21.615, df = 4, p-value = 0.000239

## alternative hypothesis: one model is inconsistent

#p-value = 0.000239 and thus below 0.05, thus FE is chosen
```

3.2.1.4 Independence of Errors

```
#Durbin Watson Test

#D-W values as close as possible to 2 but at least in between 1.5 and 2.5

durbinWatsonTest(model_5)

##  lag Autocorrelation D-W Statistic p-value

##    1      0.00698574      1.986015   0.302

##  Alternative hypothesis: rho != 0

durbinWatsonTest(model_6)

##  lag Autocorrelation D-W Statistic p-value

##    1     0.007103629       1.98578   0.326

##  Alternative hypothesis: rho != 0

pdwtest(model_7)

##

##  Durbin-Watson test for serial correlation in panel models

##

## data:  retweet_count ~ misogynism_score + URL_dummy + hashtags_dummy +
followers_count

## DW = 1.9349, p-value = 2.953e-06
## alternative hypothesis: serial correlation in idiosyncratic errors
pdwtest(model_8)
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  retweet_count ~ mis_binary + URL_dummy + hashtags_dummy + followers_count
## DW = 1.9347, p-value = 2.744e-06
## alternative hypothesis: serial correlation in idiosyncratic errors
```

4. Additional Analyses

4.1 Determining Percentage of deleted Tweets

```
#check 4 weeks after end of data collection how many tweets have been deleted


#create and save status_id list
id_list<-sample %>% select(status_id)


statuses <- id_list$status_id
saveRDS(id_list,"Tweet_ID")
write.table(id_list,file="Tweet_ID.txt",sep=";")
#DON'T RUN
#lookup tweets data for given statuses
collection2.0 <- lookup_statuses(statuses)
saveRDS(collection2.0,"collection2.0")
#instead read data in
collection2.0<-readRDS("collection2.0") #16582
19359-16582
## [1] 2777
#2777 are 14.34% from 19359
```

4.2 Mean Toxicity deleted Tweets

```
#data.frame with the deleted tweets
difference <- setdiff(sample$status_id, collection2.0$status_id)
difference<-as.data.frame(difference)
colnames(difference)<-"text_id"
```

```
#merging deleted tweets with final data
tox_average_diff<-merge(difference, final, by="text_id")


#average toxicity score of deleted tweets
mean(tox_average_diff$SEVERE_TOXICITY,na.rm=T)
```

```
## [1] 0.06947278
```

```
#add column with status deleted/not deleted
final$deleted<- final$text_id %in% difference$text_id
table(final$deleted)
```

```
##
## FALSE  TRUE
## 16582  2777
```

```
describeBy(final$SEVERE_TOXICITY,final$deleted)
```

```
##
##  Descriptive statistics by group
## group: FALSE
##    vars     n mean  sd median trimmed  mad min  max range skew kurtosis
se
## X1    1 16126 0.05 0.1   0.01    0.02 0.01   0 0.93  0.93 3.02      9.7
0
## --------------------------------------------------------------
## group: TRUE
##    vars    n mean   sd median trimmed  mad min  max range skew kurtosis
se
## X1    1 2700 0.07 0.12   0.01    0.04 0.01   0 0.81  0.81 2.39     5.42
0
```

```
#sd is similar = t-test for similar variances
#mean 0.05 and 0.07


#t-test/ mean comparison


t.test(final$SEVERE_TOXICITY~final$deleted, var.equal=T,alternative = "less
")
```

```
##
##  Two Sample t-test
##
## data:  final$SEVERE_TOXICITY by final$deleted
## t = -9.6743, df = 18824, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and
group TRUE is less than 0
```

```
## 95 percent confidence interval:
##         -Inf -0.01705758
## sample estimates:
## mean in group FALSE  mean in group TRUE 
##          0.04892072          0.06947278
#p < 0.01 = toxicity mean overall group is significantly lower than mean de
leted group
```

4.3 Mean Misogynism deleted Tweets

```
#merging deleted tweets with final data
mis_average_diff<-merge(difference, final_2, by="text_id")


#average toxicity score of deleted Tweets
mean(mis_average_diff$misogynism_score,na.rm=T)
## [1] 0.01280063
#add column with status deleted/not deleted
final_2$deleted<- final_2$text_id %in% difference$text_id
table(final_2$deleted)
## 
## FALSE  TRUE 
## 16582  2777
describeBy(final_2$misogynism_score,final_2$deleted)
## 
##  Descriptive statistics by group 
## group: FALSE
##    vars     n mean   sd median trimmed mad min max range skew kurtosis s
e
## X1    1 16582 0.01 0.03      0       0   0   0 0.5   0.5 4.57    27.33 
0
## ------------------------------------------------------------ 
## group: TRUE
##    vars    n mean   sd median trimmed mad min  max range skew kurtosis s
e
## X1    1 2777 0.01 0.04      0       0   0   0 0.33  0.33  4.1    20.41 
0
#sd is similar = t-test for similar variances
#mean 0.01010341 and 0.01270091


#t-test/ mean comparison
```

```
t.test(final_2$misogynism_score~final_2$deleted, var.equal=T,alternative =
"less")
##
##  Two Sample t-test
##
## data:  final_2$misogynism_score by final_2$deleted
## t = -3.6043, df = 19357, p-value = 0.0001569
## alternative hypothesis: true difference in means between group FALSE and
group TRUE is less than 0
## 95 percent confidence interval:
##          -Inf -0.001381719
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          0.01025895          0.01280063
##p < 0.01 = misogynism mean overall group is significantly lower than mean
deleted group
```

THE END