Identifying Different Layers of Online Misogyny

Wienke Strathern,[1] **Jürgen Pfeffer**[1]

[1]School of Social Sciences and Technology, Technical University of Munich
wienke.strathern@tum.de, juergen.pfeffer@tum.de

Abstract

Social media has become an everyday means of interaction and information sharing on the Internet. However, posts on social networks are often aggressive and toxic, especially when the topic is controversial or politically charged. Radicalization, extreme speech, and in particular online misogyny against women in the public eye have become alarmingly negative features of online discussions. The present study proposes a methodological approach to contribute to ongoing discussions about the multiple ways in which women, their experiences, and their choices are attacked in polarized social media responses. Based on a review of theories on and detection methods for misogyny, we present a classification scheme that incorporates eleven different explicit as well as implicit layers of online misogyny. We also apply our classes to a case study related to online aggression against Amber Heard in the context of the allegations of domestic violence she made against Johnny Depp. We finally evaluate the reliability of Google's Perspective API—a standard for detecting toxic language—for determining gender discrimination as toxicity. We show that a large part of online misogyny, especially when verbalized without expletive terms but instead more implicitly is not captured automatically.

1 Introduction

In May 2016 actress, model, and activist Amber Heard went public and accused her then-husband, actor Johnny Depp, of intimate partner violence. She described a turbulent relationship and reported that "Johnny verbally and physically abused me throughout our relationship"[1]. She publicly posted a picture of injuries and filed for divorce. This sparked a firestorm on social media and online news sites, with commentators offering wildly differing opinions as to what happened and who was to blame. Of course, it is not possible for an outsider to know exactly what happened in this incident or what the dynamics were in the relationship. However, many were quick to make accusations and blame one or the other.

In recent years more attention has been paid to the role of women in society, unfortunately also because of cases of real hatred against them.[2] In accordance with the Pew Research Center report on online harassment (Vogels 2021), women and men are similarly often abused or threatened online. However, women are more likely than men to report being sexually harassed (16% vs. 5%) or stalked (13% vs. 9%) online. Young women are particularly often affected by sexual harassment on the Internet—33% of women under 35 say they have been sexually harassed online. With the constant growth of social media and microblogging platforms, hatred of women is becoming more prevalent, creating numerous examples of how misogyny can spread almost uncontrolled (Jane 2017b; Ging and Siapera 2018, 2019).

Misogyny refers to hatred or prejudice against women and is manifested linguistically through various means, such as marginalization, bias, animosity, intimidation or violence, and objectification (Fersini, Rosso, and Anzovino 2018; Anzovino, Fersini, and Rosso 2018). A study reveals the sheer scale and nature of online abuse faced by women and provides a resource to researchers and engineers interested in exploring the potential of machine learning in content moderation.[3] In order to handle hateful content and protect people, automated systems are being used extensively to identify potentially problematic content. But a series of Failure-to-Act reports uncovers the dark side of social media platforms, more often experienced by women who are active on social media: "how harassment, violent threats, image-based sexual abuse can be sent by strangers, at any time and in large volumes, directly into your DMs without consent and platforms do nothing to stop it"[4]. Machine learning algorithms are deployed to scan content and flag it for human moderators. For instance, the Perspective API developed by Google Jigsaw was used to flag potentially toxic content for review on Wikipedia and in the New York Times comments section.[5] One challenge is to capture the linguistic specifics of hate speech, polarizing and offensive statements. Udupa observed that users of online social media platforms have managed to bypass automatic hate speech detection methods by using creative indirect forms of linguistic expression. According to Strathern et al. alternative methods to recog-

[1]https://www.chicagotribune.com/entertainment/ct-johnny-depp-amber-heard-statement-20160531- story.html

[2]https://onlineviolencewomen.eiu.com/

[3]https://decoders.amnesty.org/projects/troll-patrol/findings

[4]https://counterhate.com/research/hidden-hate/

[5]https://perspectiveapi.com/case-studies/

nize moral slurs could be successfully implemented.

Since hate is expressed in many different ways, automated methods can lack context sensitivity when determining implicit hate. To shed light on this discrepancy, we first examine which scientific theories and methods deal with the topic of misogyny. In the second step, we examine more closely how, based on theory and empirical work, classes of misogyny are built according to which content of hate speech can be assigned. In this, we assume that, in addition to a large amount of explicit hate speech, there is also a significant proportion of implicit misogynistic hate. Consequently, another goal of our study is to examine how well automated approaches to detect toxic language can identify misogyny. We collected 240,000 tweets from 2019–2021 containing the tweet handle @realamberheard and selected the top 5,000 most retweeted tweets to label and score them according to the classes identified in the literature. We then had these 5000 tweets analyzed by the Google Perspective API toxicity metric. A major outcome of this study is that online misogyny cannot be satisfactorily identified with this automated toxicity identification tool.

2 Review of Theories and Methods on Misogyny

Our study is motivated by work dealing with a) misogyny, its modeling and detection, b) the classification of hate speech and c) the verification of hate speech detectors.

2.1 Misogyny

As per Allen, there is no universally accepted definition of misogyny. When studying online anti-feminist language, different terms have been used, including "gender hate speech" (Jane 2015), "gender trolling" (Mantilla 2013), "cyber harassment" (Citron 2014), "technological violence" (Ostini and Hopkins 2015), "e-bile," and "gender cyber hatred" (Jane 2017a), as summarized by McGuirk. According to Code, misogyny can manifest in sexual and physical violence, exclusion, promotion of patriarchy, belittlement, or marginalization of women. Zuckerberg has supplemented this framework with specific forms of online misogyny. Jane identifies technological determinism as a paradigm of flaming. However, research on flaming does not show that online abuse is gender-specific (Lee 2016). In contrast, Herring and Martinson found that the "gendered nature" of online abuse messages and hate speech is significant when examining gender differences in communication styles. Online misogyny can have real-world consequences that require further investigation. Citron and Norton hypothesize that the gendered nature of online harassment and digital abuse is critical to women's online identity. Megarry has studied the psychological effects of online misogyny, including pseudonymous involvement and pullback, which limit women's online engagement.

The case of Amber Heard was the subject of a study by Whiting et al.. They conducted their study from a psychological perspective on the subject of domestic violence. The authors examined the commenting behavior of users on various social media platforms. To better understand typical types of social media reactions to allegations of domestic violence, the authors performed a content analysis on Facebook. Five main categories were extracted, namely victim blaming, perpetrator blaming, couple blaming, withholding judgment, and mixed reactions to the process. The respective main topics also contain subtopics on reactions to the allegations.

2.2 Modeling Misogyny

Determining and classifying misogyny in comments is a major challenge for humans and computers. There are various definitions and approaches to modeling this complex social and linguistic phenomenon. Fersini, Rosso, and Anzovino developed a machine learning classification approach to model misogyny. The main categories are based on gender studies theory and contain classes that are used to determine comments. The classes are: stereotyping and objectification, dominance, derailment, sexual harassment, threats of violence, and discrediting. The categorization starts after an a priori distinction of whether a tweet is classified as misogynistic or not. In a study by Farrell et al. a misogyny model was developed to examine the flow of extreme language in online communities on Reddit. Based on feminist language criticism, the author created nine lexicons that capture specific misogyny rhetoric (physical violence, sexual violence, hostility, patriarchy, stoicism, racism, homophobia, disparagement, and inverted narrative), and used these lexicons to examine how language evolves within and between misogynist groups. Recent work by Guest et al. presents a hierarchical taxonomy for online misogyny and an expert-labeled data set that allows automatic classification of misogyny content. The taxonomy consists of misogynistic content, broken down into misogynistic pejoratives and treatment, misogynistic disparagement, and gendered personal attacks.

2.3 Detecting Online Misogyny

In addition to modeling misogyny and detecting hate speech, we find studies examining how politically and socially active women are treated in current public debates. To gain insight into gender discrimination, various automated methods are used. In a study by Rheault, Rayment, and Musulan, the authors applied machine learning models to predict rudeness directed at Canadian politicians and US senators on Twitter. In particular, they test whether women in politics are more affected by online abuse, as recent media reports suggest. Another article by Beltran et al. examined gender insults towards Spanish female politicians. In an analysis of tweets written by citizens, the authors found evidence of gender slurs and note that mentions of appearance and infantilizing words are disproportionately common in texts addressing female politicians in Spain. The results show how citizens treat politicians differently depending on their gender. Fuchs and Schäfer presented the results of an exploratory analysis of misogynistic and sexist hate speech and abuse against female politicians on Twitter, using computer-assisted corpus linguistic tools and methods, supplemented by a qualitative in-depth study of abuse by four prominent female politicians in Japan. Studies suggest that voters evaluate candi-

dates from the perspective of gender stereotypes and test how this affects attitudes and voting behavior (Bauer 2015; Ditonto, Hamilton, and Redlawsk 2014; Herrnson, Lay, and Stokes 2003; Lawless 2015).

2.4 Hate Speech Classification

The annotation of hate speech is important for automated classification tasks. The classification scheme and its underlying assumptions are crucial for annotation. There are different approaches to this process such as predefined word lists or more complex models. One of the main difficulties is the definition of hate speech and its interpretation and therefore correct application. Recently, the Gab Hate Corpus was published (Kennedy et al. 2022), which uses a specially developed coding typology for annotating hateful comments. It was developed based on a synthesis of hate speech definitions drawn from legal precedents, hate speech coding classifications, and definitions from sociology and psychology. Moreover, the system includes a hierarchical clustering technique to identify dehumanizing and aggressive language, markers for targeted groups, and rhetorical features. Ben-David and Fernández researched the circulation of explicit hate speech and subtle forms of discrimination on Facebook. They contend that hate speech and discrimination cannot solely be attributed to the users' intentions and behaviors. It is also influenced by the interplay between the platform's policies, technological capabilities, and communicative practices of its users. The difficult task of capturing implicit and explicit statements was addressed in a study by Gao, Kuppersmith, and Huang. The writers suggested a technique for identifying online hate speech that employs a weakly supervised two-path bootstrapping method. This approach utilizes extensive unmarked data to overcome some constraints of supervised hate speech classification procedures, including dataset bias and the prohibitive expense of annotation. The implicitness of linguistic statements is also the subject of a work by Frenda, Patti, and Rosso. The authors proposed a number of statistical and computational analyses that support reflections on indirect propositions that focus on the creative and cognitive aspects of implicitness. In a more recent work by ElSherief et al., implicit statements were used for machine learning tasks to introduce a theoretically based taxonomy of hate speech. The research conducted by Wiegand, Ruppenhofer, and Eder focuses on identifying implicitly abusive language, meaning language that conveys abusive intent without using explicitly abusive words. Their position paper outlines the challenges in learning implicit abuse due to the limitations of current datasets and proposes changes in the dataset design to overcome these obstacles.

2.5 Bypassing Hate Speech Detection

Tricking or recalibrating automated methods results from the observation that the underlying assumptions of common machine methods do not adequately define group-specific hatred. That is, there seems to be a discrepancy between methods for operationalization tasks and the complexity of social processes. Against this background there are ways to trick hate speech detection methods or to test them for their measurement accuracy and validation. Both, cultural and associated linguistic peculiarities are thus taken into account. There are studies that try to capture culture- and language-specific hatred, which machines have difficulty recognizing. Zannettou et al. focused on examining the spread of antisemitic content. The authors carried out a large-scale quantitative analysis to discover abnormalities in language use. The results show that there are several distinct facets of antisemitic language, ranging from slurs to conspiracy theories, drawing on biblical literature and narratives expressed differently in the language. In this context, antisemitism is considered as a manifestation of hate speech, and the writers devised a technique to address it. Another investigation by Gröndahl et al. examined the efficacy of previously proposed models and datasets for categorizing hate speech. The findings revealed that none of the pre-existing models achieved satisfactory results when tested on a different dataset. The authors assert that the characteristics indicative of hate speech are not consistent across different datasets. The results show that the definitions of hate speech do not seem to be consistent and that they need further differentiation and context sensitivity. Another study by Hiruncharoenvate, Lin, and Gilbert examined ways to circumvent the observation of the state in the Chinese language, which suppresses free speech. In China, political activists use homophones (two words that are written differently and have different meaning but sound the same, e.g., brake/break) of censored keywords to avoid detection by keyword-matching algorithms. The authors claim that it is possible to expand this idea in a way that makes them difficult to counteract. One result of this work is to mathematically (and almost optimally) change the content of a post by replacing censored keywords with homophones. So, by tricking the system with linguistic creativity, they bypass the derived rules for automatic speech recognition on Weibo.

3 Overview of Misogyny Classes from the Literature

Based on the theories and methods discussed above, we have developed a classification scheme for online misogyny that covers most of the aspects discussed in the related literature. These classes include explicit and implicit misogynistic language and are presented in the following. Some of these classes are close to each other in their definitions and are not always easy to distinguish. The case study in the second part of this article will show that they significantly overlap when used for coding real-world messages. The goal of identifying misogyny classes was not to identify unambiguous definitions, but to cover a wide variety of aspects of hate against women.

3.1 Explicit Misogyny

In explicit misogynistic statements users openly attack, insult, or even threaten a woman (Waseem et al. 2017; Gao, Kuppersmith, and Huang 2017). Based on the literature presented above, we have identified the following four subcategories of explicit misogyny.

Call for action/violence. This class implies verbal

threads that intend to punish a target physically. Statements in which users call for deletion, prison, boycott, or sending the target to a psychiatric institution (Fersini, Rosso, and Anzovino 2018).

Personal insult, denigration. Personal insults and denigration intended to cause harm to a target verbally. Statements containing harmful wishes, demeaning, threatening, denigrating, inciting, defaming, use of slur words (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

Gendered personal attack. Gendered personal attacks refer to stereotypes of women. Verbal (misogynistic) attacks draw on these stereotypes. Statements that contain misogynistic speech and swearwords, revenge porn, or are sexually motivated because the target is a woman (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

Weakness of character, intellectual inferiority. Making negative judgments of a woman's moral and intellectual worth using explicit slur words. Statements that call a woman controlling, psychotic, a liar, hypocritical, narcissistic, or manipulative (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

3.2 Implicit Misogyny

Implicit statements of misogyny include cynicism and sarcasm, skepticism and distrust, insinuation, accusations, speculation and questioning of credibility, a demonstration of power, and taking a position (Waseem et al. 2017; Gao, Kuppersmith, and Huang 2017; ElSherief et al. 2021; Frenda, Patti, and Rosso 2022).

Cynicism, sarcasm. Cynicism and sarcasm represent a very derogatory attitude of a person towards others. It is expressed in an indirect form and is spiteful and bitter. Statements in which in a subliminal way, a rejecting attitude is shown (Whiting et al. 2019).

Skeptical attitude, distrust. That includes "facts" or other details to undermine a woman's account. Doubtfulness about a woman's claims or accusations. Questions whether the target had lied before and therefore cannot be trusted (Whiting et al. 2019).

Imputation. Imputation is understood as the assumption that the target behavior is motivated by flawed motivations. That includes statements that show a moral judgment, and comments where a woman is described as revenge-seeking, vindictive, attention-seeking, monetarily driven (Whiting et al. 2019).

Allegation. The category implies actions in which the evidence and allegations are challenged suggesting intentionally motivated actions. Statements of users that offer facts that refute a woman's account in spite of evidence (Whiting et al. 2019).

Speculation, denying credibility. This category includes an investigative-style attitude. Speculations and doubts about the target's behavior. In users' comments on the case, e.g., of domestic violence and its severity, we find claims about how the case might affect future reporting, users offering life stories to undermine the target's account, together with claims to personal expertise, the intent to prove something, credibility from experience, and special predictive power (Whiting et al. 2019).

Demonstration of power. The category implies a power relation between one gender and the other. Statements in which support for the man is demonstrated (Fersini, Rosso, and Anzovino 2018).

Taking position. Taking position or 'flipping the narrative' encapsulates terms and expressions that refer to the relationship between the target and the perpetrator. Statements on who is the 'perpetrator' and who is the 'victim' (Fersini, Rosso, and Anzovino 2018; Guest et al. 2021; Farrell et al. 2019).

3.3 Examples for Misogyny Classes

In order to study the prevalence of these misogynistic classes on social media, we have collected and analyzed messages addressing Amber Heard's Twitter account @realamberheard in a case study in the next section. Here, in Table 1, we show sample tweets to exemplify these classes. Since the content contains explicit hate speech and profanity, we have redacted the texts.

4 Case Study

To assess the importance of the misogyny classes presented in the article, we conducted a case study using Twitter data related to the celebrity domestic violence abuse case between Amber Heard and Johnny Depp. In the following, we describe the data and the annotation process as well as present quantitative results showing the prevalence of our explicit and implicit misogyny classes in the data.

Kennedy et al. documented that the annotation of hate speech has been shown to lead to a high level of disagreement between the annotators, see also Ross et al.. According to Mostafazadeh Davani et al. this is due to a combination of factors, including differences in understanding of the definition of hate speech, interpretation of the annotated texts, or assessment of the harm done to certain groups, i.e. inconsistent application of the definition of hate speech to different social groups.

Data. By utilizing the Twitter Academic API (Pfeffer et al. 2023) we collected 266,579 original tweets (excluding re-tweets) in January of 2022 that contained the account @realamberheard in the tweet texts. This resulted in 266,579 tweets (2019: 64,334 tweets, 2020: 117,231 tweets, 2021: 85,014 tweets). For the annotation process, we extracted 5,000 tweets that have been retweeted most often.

4.1 Annotation Process

For our case study we employed two annotators, a graduate student who is also a co-author on this paper and was instrumental in developing the misogyny classes (annotator 1), as well as an undergraduate student who was new to the topic (annotator 2). The annotators were briefed with an introduction to the topic in general and then presented with the misogyny classes. All the information presented together with coding examples was also shared in a coding manual. The manual also includes detailed descriptions of the individual coding steps and further explanations of the definition

Class	Example Tweet
Call for action/violence	Oh @realamberheard .... You ignorant witch. We ALL already know you're the guilty one here. Johnny's innocence has been proven. You're just trying to buy time, before you (hopefully) have you sit your scronny ass in a jail cell. You speak nothing but venomous lies. #JohnnyDepp
Personal insult, denigration	Seriously, how fucking sick you have to be to pull a "prank" like this on someone ? What kind of gross bitch would think pooping in people's bed is funny ? Well, apparently @realamberheard does. #JusticeForJohnnyDepp
Gendered personal attack	Not a johnny Depp fan but @realamberheard claims have more holes than swiss cheese. I dont understand females who can't make their own money and want to pocket off someone elses. It's hard to find a victim that no one sides with in todays world but I think we all call bs on AH.
Weakness of character, intellectual inferiority	Look what headline just poped up on sky news! @realamberheard you dirty little Lier! #AmberHeardIsALiar #JusticeForJohnnyDepp
Cynicism, sarcasm	@realamberheard Yes, the excitement around #JusticeLeague was huge ... definitely nothing to do with you though. Imagine being in a 4 hour movie for 5 minutes and being the most insufferable part of it.
Skeptical attitude, distrust	I just noticed the 'actor/ activist' claims in your biog @realamberheard !! Well, you certainly are an actress for real!! Only trouble is that the majority of your acting seems to be done OFF stage!! And you have set 'activism' back decades dear!! Ugh, you are some piece of work!
Imputation	@realamberheard @realamberheard Put your hand down and stop exploiting Evan's story to sway the public perception back in your favor. Don't act like you didn't break bread and hang out with Marilyn Manson for years after his relationship with ERW/ your o
Refutation	Listen bitch, I just saw a video about you demanding Depp supporter info for some legal implications!!If you want any info about me just DM me and I'll be MORE than happy to bring you upto speed!! @realamberheard I am allowed my opinion and you are scum (&u better pay my airfare!)
Speculation, denying credibility	@realamberheard You do not represent women nor survivors. I stand with Johnny Depp, Kate James, Jennifer Howell, Lily-Rose Depp, Hilda Vargas, Samantha McMillen, Katherine Kendall, Trinity Esparza and ALL THE OTHER women and men who knows your true color
Demonstration of Power	Justice for Johnny Depp outside @wbpictures studio where @realamberheard is currently filming @aquamanmovie #JohnnyDepp #JusticeForJohnnyDepp #JOHNNY #AmberHeard
Taking up a position	@realamberheard is not a victim, she is the perpetrator.

Table 1: Misogynic classes and example tweets

	Misogyny Class	Frequency	All	Misogyny
Explicit	Call for Action	681	13.6%	20.4%
(35.6%)	Personal Insult	1,649	33.0%	49.5%
	Gendered Personal Attack	730	14.6%	21.9%
	Intellectual Inferiority	1,325	26.5%	39.8%
Implicit	Cynicism/Sarcasm	367	7.3%	11.0%
(30.3%)	Skepticism/Distrust	461	9.2%	13.8%
	Imputation	556	11.1%	16.7%
	Allegation	546	10.9%	16.4%
	Speculation	305	6.1%	9.2%
	Demonstration of Power	459	9.2%	13.8%
	Taking up a Position	181	3.6%	5.4%
N			5,000	3,331

Table 2: Frequencies and proportions of misogyny classes in all 5,000 annotated tweets as well as proportions in 3,331 misogynistic tweets.

of the classes and the coding method according to the literature.

We analyzed the entire tweet at the sentence and word level, including the use of emoticons and content on the websites following URLs appearing in tweets. We looked at images, memes, or quotes, and watched linked videos. Each tweet was rated by the annotators based on all of its content. If the tweet contained statements supporting Amber Heard was neutral, or contained advertising, we annotated this tweet as *other* and ignored the tweet in the subsequent analytical steps. We used the eleven misogyny classes for annotation. After the annotation process, we created the explicit/implicit annotation from the eleven classes following the categorization described above. A single tweet could be annotated with multiple misogyny classes. If a tweet contained multiple sentences where one was implicit and one was explicit, we chose the explicit class due to the fact that a Tweet with explicit misogynistic content will be perceived as being explicit in its entirety.

Coding 11 classes with multiple overlapping definitions will lead to low levels of completely identical annotations. However, when comparing the explicit/implicit/other classes among the two annotators, the overall level of agreement between the annotators was acceptable. We can report the following values for Krippendorff's alpha (Krippendorff 2011): explicit 0.779, implicit 0.736, other 0.867.

4.2 Prevalence of the Misogyny Classes

For further analysis of this article, the annotator 1 manually compared the annotations from both annotators for all 5,000 tweets and harmonized the annotations into a single mapping of tweets to misogyny classes. The frequencies and proportions of the classes in the overall dataset as well as in the misogynistic tweets can be seen in Table 2. Shockingly, two-thirds of the most retweeted tweets addressing Amber Heard's Twitter account have been classified into explicit (35.6%) or implicit (30.3%) classes of misogyny. While explicit and implicit classes can overlap within tweets, the meta-classes explicit/implicit are mutually exclusive (see above).

5 Comparing Misogyny Classes with Google's Perspective API

Google's Perspective API is one of the standards for identifying toxic language on online platforms and is described as "the product of a collaborative research effort by Jigsaw and Google's Counter Abuse Technology team exploring machine learning as a tool for better discussions online."[6]. In this section, we will test how well the toxicity scores of this API are capable of identifying online misogyny as operationalized with our eleven classes to get an understanding of how useful these approaches can be in automatically identifying online misogyny.

We worked out the different attributes and evaluation methods of the API as the first step for comparison. In the

[6]https://www.perspectiveapi.com/research/

second step, we applied the API to the same dataset of 5,000 tweets. For each tweet, the API specifies a range of values for each of its categories. In the third step, we compared the values using statistical methods and applied network analysis to show the co-occurrence of classes and their average toxicity value reported by the Perspective API.

5.1 Attributes of Perspective API

The Perspective API predicts the perceived impact of a comment on a conversation by evaluating the comment with a set of emotional concepts known as 'attributes', namely toxicity, severe toxicity, identity attack, offense, threat, and profanity. The returned values are in the range [0.1] and are an indicator of the likelihood that something will be perceived as toxic. The higher the score, the more likely it is that the patterns in the text are similar to the patterns in comments that others have identified as toxic. The values are intended to allow developers/users to set a threshold and ignore values below that value. Values around 0.5 indicate that the model does not know if it is similar to toxic comments. The Google recommended threshold setting is 0.7. These thresholds are central to interpretation.

	Misogyny Class	**Average Toxicity**
Explicit (35.6%)	Call for Action	0.504
ø0.572	Personal Insult	0.589
	Gendered Personal Attack	0.619
	Intellectual Inferiority	0.577
Implicit (30.3%)	Cynicism/Sarcasm	0.356
ø0.493	Skepticism/Distrust	0.527
	Imputation	0.557
	Allegation	0.423
	Speculation	0.572
	Demonstration of Power	0.436
	Taking up a Position	0.581
Other (34.1%)	Marketing/PR	0.193

Table 3: Categories and Average Toxicity for Explicity and Implicity.

5.2 Measuring Toxicity for Misogyny Classes

To measure the average toxicity for the misogyny classes, we compare Google's probability score to our manual coding by summing up the codes divided by the number of tweets in each meta-class. The results show that the average toxicity score by Google for our category of explicit misogyny is 0.572. For our category of implicit misogyny, the average score by Google is 0.493. These numbers already are a strong indicator that toxicity, as identified with the Perspective API, is a poor predictor of our variable of online misogyny, and in particular of implicit hate against women. Table 3 reveals the average toxicity scores for each class. In Figure 1 we can further see the density distribution of toxicity scores for each of the meta-classes of tweets with explicit or implicit misogyny as well as others.

In the *other* sub-figure we can clearly see that there are almost no tweets that have been identified by the Perspective API as toxic that we have not also classified as misogynistic—consequently, the automated coding does not create false positives. The explicit language used for the

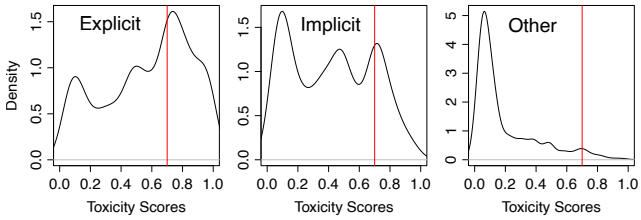

Figure 1: Distribution of toxicity scores from Google's Perspective API for tweets with explicit or implicit misogyny as well as tweets without misogynistic content.

classes that we have summarized with the meta-class *explicit* can be identified by the Perspective API to a certain degree, and the peak of the score distribution is above the standard threshold of 0.7. In other words, tweets coded with explicit misogyny contain text patterns that are similar to the patterns in comments that have been identified as toxic when the Perspective API models have been trained.

Unfortunately, the picture looks different when looking at the distribution of scores for the implicit misogyny classes. Here, the resulting toxicity scores are almost evenly distributed, having more scores with very low values than with very high values. Consequently, the tweets coded with implicit misogynistic classes do not reflect text patterns that are similar to the patterns that have been identified as toxic in the Perspective API's training data.

5.3 Co-Occurrence Network of Misogyny Classes

In addition to statistical analysis, we built a co-occurrence network that maps manual coded classes and the average toxicity scores by the Perspective API (3). Nodes represent the eleven classes and the edge value is the number of co-occurrences, i.e., the co-occurrence of classes within a tweet. The edge color is the edge value, and the node size is the proportion of the number per code divided by the number of tweets. The node color is the average toxicity value from the Perspective API where blue means low and red means high toxicity values.

In the centre, we can find the dominant four explicit classes which are identified to a certain degree as being toxic. The classes are well connected with each other. Explicit abusive statements come with similar forms of abusive language. For implicit statements, the picture looks different. In the periphery, we can find the seven classes of our meta-class implicit. Implicit misogynistic statements occur more with various forms of explicit abusive language and less among each other. In many cases, something is said implicitly, but it co-occurs with an explicit abusive statement. As mentioned above, we decided to code a tweet as explicit if both classes occurred. But the network analysis reveals the co-occurrence of explicit and implicit abusive language against women within one statement. It offers a more qualitative comparison of stereotypical hating: statements that contain a demonstration of power are associated with inferiority and insults. A skeptical attitude is associated with abusive terms of inferiority, imputation, gendered

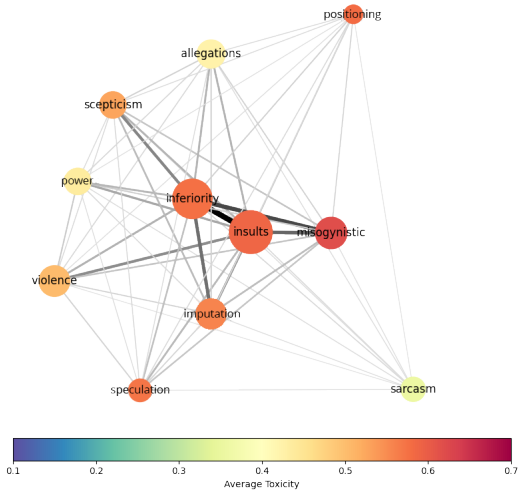

Figure 2: Co-Occurrences of Categories within a Tweet

personal attacks, and insults. Statements of speculation and doubt are associated with sarcastic and gender-attacking language. Despite the proximity of all classes, the network reveals a distinction between explicit and implicit misogyny.

5.4 Interpretation and Conclusion

We asked how well an automated approach like Google's Perspective API performs in detecting misogyny. Based on our study, two things become apparent: Google's text model does recognize explicit misogyny in the text patterns as toxic. However, the model does not recognize implicit misogyny in text patterns as toxic. The interpretation of the following tweets underlines the challenges of detecting and understanding implicit/indirect hate: "@realamberheard It's the way you think that posing this is going to change public perception of you. We heard what you did in your own words. A failure in the system isn't uncommon, so thank you for proving that male victims will never be taken seriously." A user recapitulates what has happened, draws conclusions for men, and thanks the target person for that in a very calm manner. But reading the tweet with contextual knowledge makes one understand that the thankful gesture is a cynical one. No keyword of hate can be found here; the words are all positive, but the underlying assumption is an accusation against Amber Heard and against her gender. None of the scores indicates harm in this tweet: Toxicity: 0.28, Severe Toxicity: 0.17, Identity Attack: 0.26, Offense: 0.07, Threat: 0.21 and Profanity: 0.14.

In another tweet, a user comments on what has happened and concludes that this behavior is not acceptable. The tweet contains a link to a screenshot in which impressions of what happened are reflected. Again, there is no harmful word, it all sounds positive in isolation, but clearly implies that this user is rejecting the behavior of the woman and at the same time accusing her of what she has done: "@realamberheard I had to translate to really understand where you're coming from. And no I wouldn't encourage my daughter or sister to do what you did (URL redacted)". But here as well, the scoring is very low. Toxicity: 0.20, Severe Toxicity: 0.12, Identity Attack: 0.11, Offense: 0.07, Threat: 0.16 and Profanity: 0.14.

The following example can exemplify how the toxicity score can be influenced by a single word that is interpreted as negative, even though the tweet could be interpreted as being funny: "@realamberheard @USNatArchives She will forever be known as the lady who pooped on Johnny Depp's bed." Toxicity: 0.69, Severe Toxicity: 0.15, Identity Attack: 0.74, Offense: 0.65, Threat: 0.34, and Profanity: 0.74.

There may be several reasons for this discrepancy to detect misogyny. One reason could be that there was no misogynistic content in the training texts for the human annotators. Or misogyny was never defined as an annotation class, hence, annotator could not label it. Annotators could not be informed / trained on the topic of misogyny and, therefore, could not recognize and annotate it in the texts. Although we do not know how the data sets were constructed and the model trained, we can summarize that Google's Perspective API struggles with identifying text patterns containing implicit misogynistic statements.

6 Discussion

In this manuscript, we have presented a classification scheme that incorporates 11 classes of misogyny and have described a data set that contains misogynistic content labels from Twitter. We have also provided a detailed coding book and a data set with all of the labels. The data set benefits from a detailed classification scheme based on the existing literature on online misogyny. The involvement of trained annotators and an adjudication process also ensures the quality of the labels.

We applied the classification scheme to a case related to online aggression against Amber Heard in the context of her allegations of domestic violence against Johnny Depp. For 5,000 tweets, we identified online misogyny operationalized with our eleven classes for two-thirds of the tweets, one-third as explicit misogyny, and one-third as implicit misogyny. Finally, we evaluated the reliability of Google's Perspective API for determining implicit misogyny and found that this approach can identify explicit misogyny to a certain extent, but fails with identifying implicit misogyny.

Ethical considerations and limitation. Ethical considerations must be taken into account with regard to the training and supervision of the annotator. An undergraduate student was the annotator, who underwent two steps: first, reading the typology and coding manual, and second, conducting a test on about 50 messages that had already been annotated and validated by one of the authors. Kennedy et al. pointed out the pressing concern that annotators may experience trauma or similar negative effects such as desensitization when annotating hate speech. On the basis of our own annotation experiences, we would like to highlight these thoughts. While no studies have investigated the repercussions of continuous, daily exposure to hate speech on human moderators, existing evidence suggests that being exposed to violent language and images online can adversely impact mental health, as demonstrated by Kwan et al.. We

also provided the annotator with Kennedy's suggested written guide [7] to help detect changes in cognition and avoid secondary trauma. It advises the user to take breaks and not imagine traumatic situations. The annotator was instructed to remain in communication with the study's author if she experiences any symptoms of PTSD, which are also outlined in the guide. The guide aims to normalize negative emotions resulting from work, offer education regarding trauma, identify signs of traumatic stress, and establish a support system as a preventative measure against secondary traumatic stress.

A limitation of this study is the fact that we do not know whether the Perspective API's text models contained misogynistic content and we do not know whether the data sets contained implicit/indirect forms of hate. Furthermore, we do not know whether the annotators were informed or trained on the topic of misogyny or implicit/indirect forms of hate. However, our results show that there may be a lack of information on misogyny according to existing definitions.

Google's Perspective API is a prominent tool for recognizing hate speech that uses machine learning to reduce toxicity, which is an important step towards addressing the challenge of online abuse and harassment. The API calculates the probability that a comment is perceived as toxic, reflecting Google's ambitious goal to prevent online toxicity and protect marginalized voices in conversation: "Toxicity online poses a serious challenge for platforms and publishers. Online abuse and harassment silence important voices in conversation, forcing already marginalized people offline".[8]

To evaluate the tool's effectiveness, we believe it is legitimate to directly compare it to misogyny, which represents abuse and hate according to the definition of toxicity. Given that misogyny is often subtle and has various layers, it is necessary to observe and document specific situations to collect as many characteristics as possible. We hope that by taking this approach, we can encourage the developers to adjust the tool's performance and better address the issue of online toxicity.

Implications and Future Work. Given real-world online aggression against women, it is probable that Google's toxicity model would not identify it. Thus, a huge fraction of implicit misogyny texts would stay be left in place and would not be deleted or otherwise acted upon. Misogynous behavior and target classification still remain a very challenging problem. One approach may be to create lexicons capturing specific misogynistic rhetoric and improve annotation scheme. Another challenge is to capture the peculiarities of implicity or indirect forms of hate in language. Language is very context-sensitive, and a negative tone can be expressed without a clear negative key word. Moreover, implicit sentences depend decisively on the non-linguistic accompanying signals. With our work, we would like to enhance existing research on investigating linguistic distinction between implicit and group-specific hate rhetoric. Furthermore, as we have seen from the network perspective, aside from the technical solution questions arise on how and why these different sub-classes are closely connected. From a gender perspective, we ask why are these stereotypes so consistent over time?

Given the still increasing number of users and posts in social media, automated annotation based on machine learning is inevitable. There is no other way to handle the vast volume of text. At the same time, it becomes apparent that the proportion of aggressive misogynistic speech is increasing sharply. An assessment and, if necessary, the deletion of unacceptable statements is imperative for the protection of people. Especially with regard to women, their protection is of immense importance to enable participation in public discourse and avoid withdrawing because of fear of being attacked or marginalized. However, the key to better handling the problem is to better understand the phenomenon of misogyny.

[7] https://www.apa.org/ptsd-guideline/ptsd.pdf

[8] https://perspectiveapi.com/

7 Acknowledgments

The author(s) gratefully acknowledge the financial support from the Technical University of Munich - Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

The labeled dataset, the classification schema and the coding manual are publicly available: https://strathern.de/publications.

References

Allen, A. 2021. Feminist Perspectives on Power. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Anzovino, M.; Fersini, E.; and Rosso, P. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *In International Conference on Applications of Natural Language to Information Systems*, 57–64.

Bauer, N. M. 2015. Emotional, Sensitive, and Unfit for Office? Gender Stereotype Activation and Support Female Candidates. *Political Psychology*, 36.

Beltran, J.; Gallego, A.; Huidobro, A.; Romero, E.; and Padró, L. 2021. Male and female politicians on Twitter: A machine learning approach. *European Journal of Political Research*, 60: 239–251.

Ben-David, A.; and Fernández, A. M. 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal of Communication*, 10.

Citron, D. K. 2014. Hate Crimes in Cyberspace - Introduction. *Harvard University Press*.

Citron, D. K.; and Norton, H. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, 91.

Code, L. 2003. *Encyclopedia of Feminist Theories*. Routledge.

Ditonto, T. M.; Hamilton, A. J.; and Redlawsk, D. P. 2014. Gender Stereotypes, Information Search, and Voting Behavior in Political Campaigns. *Political Behavior*, 36: 335–358.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Association for Computational Linguistics.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*.

Fersini, E.; Rosso, P.; and Anzovino, M. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *EVALITA@CLiC-it*.

Frenda, S.; Patti, V.; and Rosso, P. 2022. Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*. Publisher: Cambridge University Press.

Fuchs, T.; and Schäfer, F. 2021. Normalizing misogyny: hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum*, 33(4).

Gao, L.; Kuppersmith, A.; and Huang, R. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

Ging, D.; and Siapera, E. 2018. Special issue on online misogyny. *Feminist Media Studies*, 18.

Ging, D.; and Siapera, E., eds. 2019. *Gender Hate Online: Understanding the New Anti-Feminism*. Springer International Publishing.

Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All You Need is "Love": Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.

Guest, E.; Vidgen, B.; Mittos, A.; Sastry, N.; Tyson, G.; and Margetts, H. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Association for Computational Linguistics*, 1336–1350.

Herring, S. C.; and Martinson, A. 2004. Assessing Gender Authenticity in Computer-Mediated Language Use: Evidence From an Identity Game. *Journal of Language and Social Psychology*, 23(4). Publisher: SAGE Publications Inc.

Herrnson, P. S.; Lay, J. C.; and Stokes, A. K. 2003. Women Running "as Women": Candidate Gender, Campaign Issues, and Voter-Targeting Strategies. *The Journal of Politics*, 65(1).

Hiruncharoenvate, C.; Lin, Z.; and Gilbert, E. 2015. Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions. *Proceedings of the International AAAI Conference on Web and Social Media*, 9: 150–158. Number: 1.

Jane, E. 2017a. Gendered cyberhate: A new digital divide? In *Theorizing Digital Divides*. Routledge.

Jane, E. 2017b. *Misogyny Online: A Short and (Brutish) History*. Sage Publications.

Jane, E. A. 2015. Flaming? What flaming? The pitfalls and potentials of researching online hostility. *Ethics and Information Technology*, 17: 65–87.

Kennedy, B.; Atari, M.; Davani, A. M.; Yeh, L.; Omrani, A.; Kim, Y.; Coombs, K.; Havaldar, S.; Portillo-Wightman, G.; Gonzalez, E.; Hoover, J.; Azatian, A.; Hussain, A.; Lara, A.; Cardenas, G.; Omary, A.; Park, C.; Wang, X.; Wijaya, C.; Zhang, Y.; Meyerowitz, B.; and Dehghani, M. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56.

Krippendorff, K. 2011. *Computing krippendorff's alpha-reliability*. Retrieved from: https://repository.upenn.edu/ascpapers/43.

Kwan, I.; Dickson, K.; Richardson, M.; MacDowall, W.; Burchett, H.; Stansfield, C.; Brunton, G.; Sutcliffe, K.; and Thomas, J. 2020. Cyberbullying and Children and Young People's Mental Health: A Systematic Map of Systematic Reviews. *Cyberpsychology, Behavior and Social Networking*, 23(2).

Lawless, J. L. 2015. Female Candidates and Legislators. *Annual Review of Political Science*, 18.

Lee, H. 2016. Behavioral Strategies for Dealing with Flaming in An Online Forum. *The Sociological Quarterly*, 46(2): 385–403.

Mantilla, K. 2013. Gendertrolling: Misogyny Adapts to New Media. *Feminist Studies*, 39(2): 563–571.

McGuirk, O. 2021. Where Have All The Good Men Gone? An Exploration of Misogyny and Anti-Feminist Discourse in the 'Mansophere'. *Thesis, Dun Laoghaire Institute of Art, Design, and Technology*.

Megarry, J. 2014. Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47: 46–55.

Mostafazadeh Davani, A.; Atari, M.; Kennedy, B.; Havaldar, S.; and Dehghani, M. 2020. Hatred is in the Eye of the Annotator: Hate Speech Classifiers Learn Human-Like Social Stereotypes. In *31st Annual Conference of the Cognitive Science Society (CogSci)*.

Ostini, J.; and Hopkins, S. 2015. Online harassment is a form of violence. *The Conversation*, 8: 1–4. Publisher: The Conversation Media Trust.

Pfeffer, J.; Mooseder, A.; Lasser, J.; Hammer, L.; Stritzel, O.; and Garcia, D. 2023. This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Rheault, L.; Rayment, E.; and Musulan, A. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1).

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M.; Wojatzki, M.; and Zesch, T., eds., *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, 6–9.

Strathern, W.; Schoenfeld, M.; Ghawi, R.; and Pfeffer, J. 2022. Identifying Lexical Change in Negative Word-of-Mouth on Social Media. *Social Network Analysis and Mining*, 59(12).

Udupa, S. 2020. Artificial Intelligence and the Cultural Problem of Extreme Speech. *Social Science Research Council, (20 December 2020), online*.

Vogels, E. A. 2021. The State of Online Harassment. *Pew Research Center*.

Waseem, Z.; Davidson, T.; Warmsley, D.; and Weber, I. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78–84. Vancouver, BC, Canada: Association for Computational Linguistics.

Whiting, J.; Olufowote, R. D.; Cravens-Pickens, J.; and Witting, A. B. 2019. Online Blaming and Intimate Partner Violence: A Content Analysis of Social Media Comments. *The Qualitative Report*, 24.

Wiegand, M.; Ruppenhofer, J.; and Eder, E. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 576–587. Association for Computational Linguistics.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A Quantitative Approach to Understanding Online Antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zuckerberg, D. 2018. *Not All Dead White Men. Classics and Misogyny in the Digital Age*. Harvard University Press.