Annotating Online Misogyny

Philine Zeinert
IT University of Copenhagen
Denmark
phze@itu.dk

Nanna Inie
IT University of Copenhagen
Denmark
nans@itu.dk

Leon Derczynski
IT University of Copenhagen
Denmark
leod@itu.dk

Abstract

Online misogyny, a category of online abusive language, has serious and harmful social consequences. Automatic detection of misogynistic language online, while imperative, poses complicated challenges to both data gathering, data annotation, and bias mitigation, as this type of data is linguistically complex and diverse. This paper makes three contributions in this area: Firstly, we describe the detailed design of our iterative annotation process and codebook. Secondly, we present a comprehensive taxonomy of labels for annotating misogyny in natural written language, and finally, we introduce a high-quality dataset of annotated posts sampled from social media posts.

1 Introduction

Abusive language is a phenomenon with serious consequences for its victims, and misogyny is no exception. According to a 2017 report from Amnesty International, 23% of women from eight different countries have experienced online abuse or harassment at least once, and 41% of these said that on at least one occasion, these online experiences made them feel that their physical safety was threatened (Amnesty International, 2017).

Automatic detection of abusive language can help identify and report harmful accounts and acts, and allows *counter* narratives (Chung et al., 2019; Garland et al., 2020; Ziems et al., 2020). Due to the volume of online text and the mental impact on humans who are employed to moderate online abusive language - moderators of abusive online content have been shown to develop serious PTSD and depressive symptoms (Casey Newton, 2020) - it is urgent to develop systems to automate the detection and moderation of online abusive language. Automatic detection, however, presents significant challenges (Vidgen et al., 2019).

Abusive language is linguistically diverse (Vidgen and Derczynski, 2020), both explicitly, in the form of swear words or profanities; implicitly, in the form of sarcasm or humor (Waseem et al., 2017); and subtly, in the form of attitudes and opinions. Recognizing distinctions between variants of misogyny is challenging for humans, let alone computers. Systems for automatic detection are usually created using labeled training data (Kiritchenko et al., 2020), hence, their performance depends on the quality and representativity of the available datasets and their labels. We currently lack transparent methods for how to create diverse datasets. When abusive language is annotated, classes are often created based on each unique dataset (a purely inductive approach), rather than taking advantage of general, established terminology from, for instance, social science or psychology (a deductive approach, building on existing research). This makes classification scores difficult to compare and apply across diverse training datasets.

This paper investigates the research question: *How might we design a comprehensive annotation process which results in high quality data for automatically detecting misogyny?* We make three novel contributions: **1.** Methodology: We describe our iterative approach to the annotation process in a transparent way which allows for a higher degree of comparability with similar research. **2.** Model: We present a taxonomy and annotation codebook grounded in previous research on automatic detection of misogyny as well as social science terminology. **3.** Dataset: We present a new, annotated corpus of Danish social media posts, Bajer,[1] annotated for misogyny, including analysis of class balance, word frequencies, Inter-Annotator Agreement (IAA), annotation errors, and classification baseline.

[1] https://github.com/phze22/Online-Misogyny-in-Danish-Bajer

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3181–3197
August 1–6, 2021.

Since research has indicated that misogyny presents differently across languages, and, likely, cultures (Anzovino et al., 2018), an additional contribution of this work is that it presents a dataset of misogyny in *Danish*, a North Germanic language, spoken by only six million people, and indeed the first work of its kind in any Scandinavian/Nordic culture to our knowledge. In Denmark an increasing proportion of people refrain from online discourse due to the harsh tone, with 68% of social media users self-excluding in 2021 (Analyse & Tal, 2021; Andersen and Langberg, 2021), making this study contextually relevant. Further, the lack of language resources available for Danish (Kirkedal et al., 2019) coupled with its lexical complexity (Bleses et al., 2008) make it an intricate research objective for natural language processing.

2 Background and related work

Abusive language is as ancient a phenomenon as written language itself. Written profanities and insults about others are found as old as graffiti on ruins from the Roman empire (Wallace, 2005). Automatic processing of abusive text is far more recent, early work including e.g. Davidson et al. (2017) and Waseem et al. (2017). Research in this field has produced both data, taxonomies, and methods for detecting and defining abuse, but there exists no objective framing for what constitutes abuse and what does not. In this work, we focus on a specific category of online abuse, namely *misogyny*.

2.1 Online misogyny and existing datasets

Misogyny can be categorised as a subbranch of hate speech and is described as *hateful content targeting women* (Waseem, 2016). The degree of toxicity depends on complicated subjective measures, for instance, the receiver's perception of the dialect of the speaker (Sap et al., 2019).

Annotating misogyny typically requires more than a binary present/absent label. Chiril et al. (2020), for instance, use three categories to classify misogyny in French: *direct* sexist content (directly addressed to a woman or a group of women), *descriptive* sexist content (describing a woman or women in general) or *reporting* sexist content (a report of a sexism experience or a denunciation of a sexist behaviour). This categorization does not, however, specify the *type* of misogyny.

Jha and Mamidi (2017) distinguish between *harsh* and *benevolvent* sexism, building on the data from the work of Waseem and Hovy (2016). While harsh sexism (hateful or negative views of women) is the more recognized type of sexism, benevolent sexism ("a subjectively positive view towards men or women"), often exemplified as a compliment using a positive stereotypical picture, is still discriminating (Glick and Fiske, 1996). Other categorisations of harassment towards women have distinguished between physical, sexual and indirect occurrences (Sharifirad and Jacovi, 2019).

Anzovino et al. (2018) classify misogyny more segregated in five subcategories: *Discredit*, *Harassment & Threats of Violence*, *Derailing*, *Stereotype & Objectification*, and *Dominance*. They also distinguish between if the abuse is active or passive towards the target. These labels appear to apply well to other languages, and quantitative representation of labels differ by language. For example, Spanish shows a stronger presence of *Dominance*, Italian of *Stereotype & Objectification*, and English of *Discredit*. As we see variance across languages, building terminology for labeling misogyny correctly is therefore a key challenge in being able to detect it automatically. Parikh et al. (2019) take a multi-label approach to categorizing posts from the "Everyday Sexism Project", where as many as 23 different categories are not mutually exclusive. The types of sexism identified in their dataset include *body shaming*, *gaslighting*, and *mansplaining*. While the categories of this work are extremely detailed and socially useful, several studies have demonstrated the challenge for human annotators to use labels that are intuitively unclear (Chatzakou et al., 2017; Vidgen et al., 2019) or closely related to each other (Founta et al., 2018).

Guest et al. (2021) suggest a novel taxonomy for misogyny labeling applied to a corpus of primarily English Reddit posts. Based on previous research, including Anzovino et al. (2018), they present the following four overarching categories of misogyny: (i) Misogynistic Pejoratives, (ii) descriptions of Misogynistic Treatment, (iii) acts of Misogynistic Derogation and (iv) Gendered Personal attacks against women.

The current work combines previous categorizations on misogyny into a taxonomy which is useful for annotation of misogyny in all languages, while being transparent about the construction of this taxonomy. Our work builds on the previous work presented in this section, continuous discussions among the annotators, and the addition of social

science terminology to create a single-label taxonomy of misogyny as identified in Danish social media posts across various platforms.

3 Methodology and dataset creation

The creation of quality datasets involves a chain of methodological decisions. In this section, we will present the rationale of creating our dataset under three headlines: Dataset, Annotation process, and Mitigating biases.

3.1 Dataset: Online misogyny in social media

Bender and Friedman (2018) present a set of *data statements* for NLP which help "alleviate issues related to exclusion and bias in language technology, lead[ing] to better precision in claims about how natural language processing research can generalize and thus better engineering results".

Data statements are a characterization of a dataset which provides context to others to understand how experimental results might generalize and what biases might be reflected in systems built on the software. We present our data statements for the dataset creation in the following:

Curation rationale: Random sampling of text often results in scarcity of examples of specifically misogynistic content (e.g. (Wulczyn et al., 2017; Founta et al., 2018)). Therefore, we used the common alternative of collecting data by using predefined keywords with a potentially high search hit (e.g. Waseem and Hovy (2016)), and identifying relevant user-profiles (e.g. (Anzovino et al., 2018)) and related topics (e.g. (Kumar et al., 2018)).

We searched for keyword (specific slurs, hashtags), that are known to occur in sexist posts. These were defined by previous work, a slur list from Reddit, and from interviews and surveys of online misogyny among women. We also searched for broader terms like "sex" or "women", which do not appear exclusively in a misogynistic context, for example in the topic search, where we gathered relevant posts and their comments from the social media pages of public media. A complete list of keywords can be found in the appendix.

Social media provides a potentially biased, but broad snapshot of online human discourse, with plenty of language and behaviours represented. Following best practice guidelines (Vidgen and Derczynski, 2020), we sampled from a language for which there are no existing annotations of the target phenomenon: Danish.

Different social media platforms attract different user groups and can exhibit domain-specific language (Karan and Šnajder, 2018). Rather than choosing one platform (existing misogyny datasets are primarily based on Twitter and Reddit (Guest et al., 2021)), we sampled from multiple platforms: Statista (2020) shows that the platform where most Danish users are present is Facebook, followed by Twitter, YouTube, Instagram and lastly, Reddit. The dataset was sampled from Twitter, Facebook and Reddit posts as plain text.

Language variety: Danish, BCP-47: da-DK.

Text characteristics: Danish colloquial web speech. Posts, comments, retweets: max. length 512, average length: 161 characters.

Speaker demographics: Social media users, age/gender/race unknown/mixed.

Speech situation: Interactive, social media discussions.

Annotator demographics: We recruited annotators aiming specifically for diversity in gender, age, occupation/ background (linguistic and ethnographic knowledge), region (spoken dialects) as well as an additional facilitator with a background in ethnography to lead initial discussions (see Table 1). Annotators were appointed as full-time employees with full standard benefits.

Gender:	6 female, 2 male (8 total)
Age:	5 $<$30; 3 $\geq$30
Ethnicity:	5 Danish: 1 Persian, 1 Arabic, 1 Polish
Study/ occupation:	Linguistics (2); Health/Software Design; Ethnography/Digital Design; Communication/Psychology; Anthropology/Broadcast Moderator; Ethnography/Climate Change; Film Artist

Table 1: Annotators/facilitator demographics
All annotators were involved during the whole project period.

3.2 Annotation process

In annotating our dataset, we built on the *MATTER* framework (Pustejovsky and Stubbs, 2012) and use the variation presented by Finlayson and Erjavec (2017) (the *MALER* framework), where the Train

& Test stages are replaced by *Leveraging* of annotations for one's particular goal, in our case the creation of a comprehensive taxonomy.

We created a set of guidelines for the annotators. The annotators were first asked to read the guidelines and individually annotate about 150 different posts, after which there was a shared discussion. After this pilot round, the volume of samples per annotator was increased and every sample labeled by 2-3 annotators. When instances were 'flagged' or annotators disagreed on them, they were discussed during weekly meetings, and misunderstandings were resolved together with the external facilitator. After round three, when reaching 7k annotated posts (Figure 2), we continued with independent annotations maintaining a 15% instance overlap between randomly picked annotator pairs.

Management of annotator disagreement is an important part of the process design. Disagreements can be solved by majority voting (Davidson et al., 2017; Wiegand et al., 2019), labeled as abuse if at least one annotator has labeled it (Golbeck et al., 2017) or by a third objective instance (Gao and Huang, 2017). Most datasets use crowdsourcing platforms or a few academic experts for annotation (Vidgen and Derczynski, 2020). Inter-annotator-agreement (IAA) and classification performance are established as two grounded evaluation measurements for annotation quality (Vidgen and Derczynski, 2020). Comparing the performance of amateur annotators (while providing guidelines) with expert annotators for sexism and racism annotation, Waseem (2016) show that the quality of amateur annotators is competitive with expert annotations when several amateurs agree. Facing the trade-off between training annotators intensely and the number of involved annotators, we continued with the trained annotators and group discussions/ individual revisions for flagged content and disagreements (Section 5.4).

3.3 Mitigating Biases

Prior work demonstrates that biases in datasets can occur through the training and selection of annotators or selection of posts to annotate (Geva et al., 2019; Wiegand et al., 2019; Sap et al., 2019; Al Kuwatly et al., 2020; Ousidhoum et al., 2020).

Selection biases: Selection biases for abusive language can be seen in the sampling of text, for instance when using keyword search (Wiegand et al., 2019), topic dependency (Ousidhoum et al., 2020), users (Wiegand et al., 2019), domain (Wiegand et al., 2019), time (Florio et al., 2020) and lack of linguistic variety (Vidgen and Derczynski, 2020).

Label biases: Label biases can be caused by, for instance, non-representative annotator selection, lack in training/domain expertise, preconceived notions, or pre-held stereotypes. These biases are treated in relation to abusive language datasets by several sources, e.g. general sampling and annotators biases (Waseem, 2016; Al Kuwatly et al., 2020), biases towards minority identity mentions based for example on gender or race (Davidson et al., 2017; Dixon et al., 2018; Park et al., 2018; Davidson et al., 2019), and political annotator biases (Wich et al., 2020). Other qualitative biases comprise, for instance, demographic bias, over-generalization, topic exposure as social biases (Hovy and Spruit, 2016).

Systematic measurement of biases in datasets remains an open research problem. Friedman and Nissenbaum (1996) discuss "freedom from biases" as an ideal for good computer systems, and state that methods applied during data creation influence the quality of the resulting dataset quality with which systems are later trained. Shah et al. (2020) showed that half of biases are caused by the methodology design, and presented a first approach of classifying a broad range of predictive biases under one umbrella in NLP.

We applied several measures to mitigate biases occurring through the annotation design and execution: First, we selected labels grounded in existing, peer-reviewed research from more than one field. Second, we aimed for diversity in annotator profiles in terms of age, gender, dialect, and background. Third, we recruited a facilitator with a background in ethnographic studies and provided intense annotator training. Fourth, we engaged in weekly group discussions, iteratively improving the codebook and integrating edge cases. Fifth, the selection of platforms from which we sampled data is based on local user representation in Denmark, rather than convenience. Sixth, diverse sampling methods for data collection reduced selection biases.

4 A taxonomy and codebook for labeling online misogyny

Good language taxonomies systematically bring together definitions and describe general principles of each definition. The purpose is categorizing

	reference	lang.	labels
Abusive Language	Zampieri et al. (2019)	da,en, gr,ar,tu	Offensive (OFF)/Not offensive (NOT) Targeted Insult (TIN)/Untargeted (UNT)/ Individual (IND)/Group (GRP)/Other (OTH)
Hate speech	Waseem and Hovy (2016)	en	Sexism, Racism
Misogyny	Anzovino et al. (2018)	en,it,es	Discredit, Stereotype, Objectification, Sexual Harassm., Dominance, Derailing
	Jha and Mamidi (2017)	en	Benevolent extension

Table 2: Established taxonomies and their use for the misogyny detection task

and mapping entities in a way that demonstrates their natural relationship, e.g. Schmidt and Wiegand (2017); Anzovino et al. (2018); Zampieri et al. (2019); Banko et al. (2020). Their application is especially clear in shared tasks, as for multilingual sexism detection against women, SemEval 2019 (Basile et al., 2019).

On one hand, it should be an aim of a taxonomy that it is easily understandable and applicable for annotators from various background and with different expertise levels. On the other hand, a taxonomy is only useful if it is also *correct and comprehensive*, i.e. a good representation of the world. Therefore, we have aimed to integrate definitions from several sources of previous research (deductive approach) as well as categories resulting from discussions of the concrete data (inductive approach).

Our taxonomy for *misogyny* is the product of (a) existing research in online abusive language and misogyny (specifically the work in Table 2), (b) a review of misogyny in the context of online platforms and online platforms in a Danish context (c) iterative adjustments during the process including discussions between the authors and annotators.

The **labeling scheme** (Figure 1) is the main structure for guidelines for the annotators, while a **codebook** ensured common understanding of the label descriptions. The codebook provided the annotators with definitions from the combined taxonomies. The descriptions were adjusted to distinguish edge-cases during the weekly discussion rounds.

The taxonomy has four levels: (1) Abusive (abusive/not abusive), (2) Target (individual/group/others/untargeted), (3) Group type (racism/misogyny/others), (4) Misogyny type (harrassment/discredit/stereotype & objectification/dominance/neosexism/benevolent). To demonstrate the relationship of misogyny to other instances of abusive language, our taxonomy embeds misogyny as a subcategory of abusive language. Misogyny is distinguished from, for instance, personal attacks, which is closer to the abusive language of *cyberbullying*. For definitions and examples from the dataset to the categories, see Appendix A.1. We build on the taxonomy suggested in Zampieri et al. (2019), which has been applied to datasets in several languages as well as in SemEval (Zampieri et al., 2020). While Parikh et al. (2019) provide a rich collection of sexism categories, multiple, overlapping labels do not fulfill the purpose of being easily understandable and applicable for annotators. The taxonomies in Anzovino et al. (2018) and Jha and Mamidi (2017) have proved their application to English, Italian and Spanish, and offer more general labels. Some labels from previous work were removed from the labeling scheme during the weekly discussions among authors and annotators, (for instance *derailing*), because no instances of them were found in the data.

4.1 Misogyny: Neosexism

During our analysis of misogyny in the Danish context (b), we became aware of the term "neosexism". Neosexism is a concept defined in Tougas et al. (1999), and presents as *the belief that women have already achieved equality, and that discrimination of women does not exist*. Neosexism is based on covert sexist beliefs, which can "go unnoticed, disappearing into the cultural norms. Those who consider themselves supporters of women's rights may maintain non-traditional gender roles, but also exhibit subtle sexist beliefs" (Martinez et al., 2010). Sexism in Denmark appear to correlate with the modern sexism scale (Skewes et al., 2019; Tougas et al., 1995; Swim et al., 1995; Campbell et al., 1997). Neosexism was added to the taxonomy before annotation began, and as we will see in the analysis section, neosexism was the most common

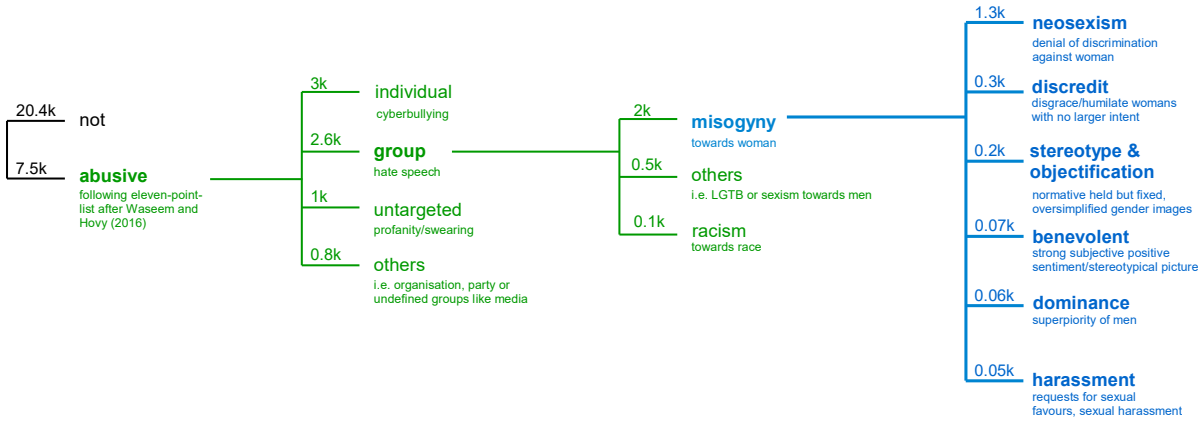


Figure 1: Labeling scheme: a taxonomy of misogyny in social media posts (blue) within abusive language categorization (green). Definitions and examples can be found in the compressed annotation codebook in Appendix A.1

form of misogyny present in our dataset (Figure 1). Here follow some examples of neosexism from our dataset:

- Resenting complaints about discrimination: *"I often feel that people have treated me better and spoken nicer to me because I was a girl, so I have a hard time taking it seriously when people think that women are so discriminated against in the Western world."*
- Questioning the existence of discrimination: *"Can you point to research showing that childbirth is the reason why mothers miss out on promotions?"*
- Presenting men as victims: *"Classic. If it's a disadvantage for women it's the fault of society. If men, then it must be their own. Sexism thrives on the feminist wing."*

Neosexism is an implicit form of misogyny, which is reflected in annotation challenges summarised in section 5.5. In prior taxonomies, instances of neosexism would most likely have been assigned to the implicit appearances of misogynistic treatment (ii) (Guest et al., 2021) – or perhaps not classified as misogyny at all. Neosexism is most closely related to the definition "disrespectful actions, suggesting or stating that women should be controlled in some way, especially by men". This definition, however, does not describe the direct denial that misogyny exists. Without a distinct and explicit neosexism category, however, these phenomena may be mixed up or even ignored.

The taxonomy follows the suggestions of Vidgen et al. (2019) for establishing unifying taxonomies in abusive language while integrating context-related occurrences. A similar idea is demonstrated in Mulki and Ghanem (2021), adding *damning* as an occurrence of misogyny in an Arabic context. While most of previous research is done in English, these language-specific findings highlight the need for taxonomies that are flexible to different contexts, i.e. they are *good representations of the world*. Lastly, from an NLP point of view, languages with less resources for training data can profit further from transfer learning with similar labels, as demonstrated in Pamungkas et al. (2020) for misogyny detection.

5 Results and Analysis

5.1 Class Balance

The final dataset contains 27.9K comments, of which 7.5K contain abusive language. Misogynistic posts comprise 7% of overall posts. *Neosexism* is by far the most frequently represented class with 1.3K tagged posts, while *Discredit* and *Stereotype & objectification* are present in 0.3K and 0.2K posts. *Benevolent*, *Dominance*, and *Harrassment* are tagged in between only 45 and 70 posts.

5.2 Domain/Sampling representation

Most posts tagged as abusive and/or containing misogyny are retrieved from searches on posts from public media profiles, see Table 3. Facebook and Twitter are equally represented, while Reddit is in the minority. Reddit posts were sampled from an available historical collection.

5.3 Word Counts

Frequencies of the words; 'kvinder' (*women*) and 'mænd' (*men*) were the highest, but these words did

samp.	domain	dis. dom	*time*	abs. in k	dis. ⊂ abus	dis. ⊂ mis
topic	**Facebook**	**48%**	***07-11/20***	12,3	51%	63%
keyw.	**Twitter**	**45%**	***08-12/20***	7,8	32%	27%
user	Twitter			3,6	8%	6%
keyw.	**Reddit**	**7%**	***02-04/19***	2,4	7%	2%
popul.	Facebook			1	2%	2%

Table 3: Distribution sampling techniques and domains
Sampling techniques: topic = posts from public media sites and comments to these posts; keyw. = keyword/hashtag-search; popul. = most interactions.

not represent strong polarities towards abusive and misogynistic content (Table 4). The word 'user' represents de-identified references to discussion participants ("@USER").

dataset	⊂ **abus**	⊂ **mis**
(kvinder, 0.29)	(kvinder, 0.34)	(kvinder, 0.41)
(user, 0.29)	(user, 0.25)	(mænd, 0.28)
(metoo, 0.25)	(mænd, 0.22)	(user, 0.18)
(mænd, 0.21)	(bare, 0.17)	(år, 0.16)
(bare, 0.16)	(metoo, 0.16)	(når, 0.15)

Table 4: Top-3 word frequencies
tf-idf scores with prior removal of special character and stop-words, notion:(token, tf-idf)

5.4 Inter-Annotator Agreement (IAA)

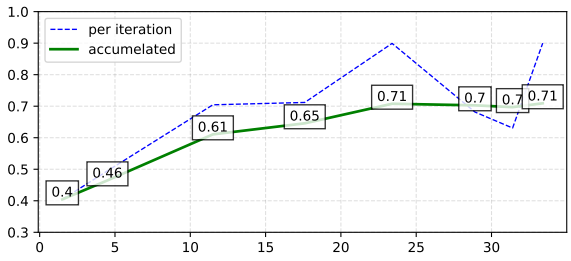


Figure 2: Inter-Annotator-Agreement
y-axis: Agreement by rel. overlap of label-sequences per sample; x-axis: Annotated data samples in k.

We measure IAA using the agreement between 3 annotators for each instance until round 3 (7k posts), and then sub-sampled data overlaps between 2 annotators. IAA is calculated through average label agreement at post level – for example if two annotators label two posts [abusive, untargeted] and [abusive, group targeted] the agreement would be 0.5. Our IAA during iterations of dataset construction ranged between 0.5 and 0.71. In the penultimate annotation round we saw a drop in agreement (Figure 2); this is attributed to a change in underlying text genre, moving to longer Reddit posts. 25% of disagreements about classifications were solved during discussions. Annotators had the opportunity to adjust their disagreed annotation in the first revision individually, which represents the remaining 75% (Table 5). The majority of disagreements were on subtask A, deciding whether the post was abusive or not.

individual corr.	group solv.	discussion round
417	169	69 (+125 pilot)

Table 5: Solved disagreements/flagged content

The final overall Fleiss' Kappa (Fleiss (1971)) for individual subtasks are: abusive/not: 0.58, targeted: 0.54, misogyny/not: 0.54. It is notable here that the dataset is significantly more skewed than prior work which upsampled to 1:1 class balances. Chance-corrected measurements are sensitive to agreement on rare categories and higher agreement is needed to reach reliability, as shown in Artstein and Poesio (2008).

5.5 Annotator disagreement analysis

Based on the discussion rounds, the following types of posts were the most challenging to annotate:

1. *Interpretation of the author's intention (irony, sarcasm, jokes, and questions)*
   E.g. ***Haha! Virksomheder i Danmark: Vi ansætter aldrig en kvinde igen...*** *(Haha! Companies in Denmark: We will never hire a woman again ...)*
   sexisme og seksuelt frisind er da vist ikke det samme? *(I don't believe sexism and sexual liberalism are the same?)*
2. *Degree of abuse: Misrepresenting the truth to harm the subject or fact*
   E.g. ***Han er en stor løgner*** *(He is a big liar)*
3. *Hashtags: Meaning and usage of hashtags in relation to the context*
   E.g. ***#nometoo***
4. *World knowledge required:*
   Du siger at Frank bruger sin magt forkert men du bruger din til at brænde så mange mænd på bålet ...
   (You say that Frank uses his power wrongly, but you use

yours to throw so many men on the fire ... - referring to a specific political topic.)

5. *Quotes: re-posting or re-tweeting a quote gives limited information about the support or denial of the author*
6. *Jargon: receiver's perception*
   I skal alle have et klap i måsen herfra *(You all get a pat on the behind from me)*

Handling these was an iterative process of raising cases for revision in the discussion rounds, formulating the issue, and providing documentation. We added the status and, where applicable, outcome from these cases to the guidelines. We also added explanations of hashtags and definitions of unclear identities, like "the media", as a company. For quotes without declaration of rejection or support, we agreed to label them as not abusive, since the motivation of re-posting is not clear.

5.6 Baseline Experiments as an indicator

Lastly, we provide a classification baseline: For misogyny and abusive language, the BERT model from Devlin et al. (2019) proved to be a robust architecture for cross-domain (Swamy et al., 2019) and cross-lingual (Pamungkas et al., 2020; Mulki and Ghanem, 2021) transfer. We use therefore multilingual BERT ('bert-base-multilingual-un cased') for general language understanding in Danish, fine-tuned on our dataset.

Model: We follow the suggested parameters from Mosbach et al. (2020) for fine-tuning (learning rate 2e-5, weight decay 0.01, AdamW optimizer without bias correction). Class imbalance is handled by weighted sampling and data split for train/test 80/20. Experiments are conducted with batch size 32 using Tesla V100 GPU.

Preprocessing: Our initial pre-processing of the unstrucutured posts included converting emojis to text, url replacement, limit @USER and punctuation occurrences and adding special tokens for upper case letters adopted from Ahn et al. (2020).

Classification: Since the effect of applying multi-task-learning might not conditionally improve performance (Mulki and Ghanem, 2021), the classification is evaluated on a subset of the dataset for each subtask (see Table 6) including all posts of the target label (e.g. misogyny) and stratified sampling of the non-target classes (e.g. for non-misogynistic: abusive and non-abusive posts) with 10k posts for each experiment. Results are reported when the model reached stabilized per class f1 scores for all classes on the test set ($\pm$ 0.01/20). The results indicate the expected challenge of accurately predicting less-represented classes and generalizing to unseen data. Analysing False Positives and False Negatives on the misogyny detection task, we cannot recognise noticeable correlations with other abusive forms and disagreements/ difficult cases from the annotation task.

subtask	epoch	f1	prec.	recall
abus/not	200	0.7650	76.43%	76.4%
target	120	0.6502	64.45%	66.2%
misog./not	**200**	**0.8549**	**85.27%**	**85.85%**
misog.*		**0.6191**		
misog.categ.	100	0.7913	77.79%	81.26%

Table 6: Baseline Evaluation: F1-scores, Precision, Recall (weighted, *except for misog., class f1-score) with mBERT

6 Discussion and reflection

Reflections on sampling We sampled from different platforms, and applied different sampling techniques. The goal was to ensure, first, a sufficient amount of misogynistic content and, secondly, mitigation of biases stemming from a uniform dataset.

Surprisingly, *topic sampling* unearthed a higher density of misogynistic content than targeted *keyword search* (Table 3). While researching platforms, we noticed the limited presence of Danish for publicly available men-dominated fora (e.g. gaming forums such as DotA2 and extremist plaftorms such as Gab (Kennedy et al., 2018)). This, as well as limitations of platform APIs caused a narrow data selection. Often, non-privileged languages can gain from cross-language transfer learning. We experimented with translating misogynistic posts from Fersini et al. (2018) to Danish, using translation services, and thereby augment the minority class data. Translation services did not provide a sampling alternative. Additionally, as discovered by Anzovino et al. (2018), misogynistic content seems to vary with culture. This makes language-specific investigations important, both for the sake of quality of automatic detection systems,

total	text corrected	label corrected	out
960	877	224	48

Table 7: Translating IberEval posts EN to DA

as well as for cultural discovery and investigation. Table 7 shows results of post-translation manual correction by annotators (all fluent in English).

Reflections on annotation process Using just seven annotators has the disadvantage that one is unlikely to achieve as broad a range of annotator profiles as, for instance, through crowdsourcing. However, during annotation and weekly discussions, we saw clear benefits from having a small annotator group with different backgrounds and intense training. While annotation quality cannot be measured by IAA alone, the time for debate clarified taxonomy items, gave thorough guidelines, and increased the likelihood of correct annotations. The latter reflects the quality of the final dataset, while the former two indicate that the taxonomy and codebook are likely useful for other researchers analysing and processing online misogyny.

6.1 A comprehensive taxonomy for misogyny

The semi-open development of the taxonomy and frequent discussions allowed the detection *neosexism* as an implicit form of misogyny. Future research in taxonomies of misogyny could consider including distinctions between active/passive misogyny, as suggested by Anzovino et al. (2018) as well as other sub-phenomena.

In the resulting dataset, we saw a strong representation of *neosexism*. Whether this is a specific cultural phenomenon for Danish, or indicative of general online behaviour, is not clear.

The use of unified taxonomies in research affords the possibility to test the codebook guidelines iteratively. We include a short version of the guidelines in the appendix; the original document consists of seventeen pages. In a feedback survey following the annotation work, most of the annotators described that during the process, they used the guidelines primarily for revision in case they felt unsure how to label the post. To make the annotation more intuitively clear for annotators, we suggest reconsidering documentation tools and their accessibility for annotators. Guidelines are crucial for handling linguistic challenges, and well-documented decisions about them serve to create comparable research on detecting online misogyny across languages and dataset.

7 Conclusion and future work

In this work, we have documented the construction of a dataset for training systems for automatic detection of online misogyny. We also present the resulting dataset of misogyny in Danish social media, Bajer, including class balance, word counts, and baseline as an indicator. This dataset is available for research purposes upon request.

The objective of this research was to explore the design of an annotation process which would result in a high quality dataset, and which was transparent and useful for other researchers.

Our approach was to recruit and train a diverse group of annotators and build a taxonomy and codebook through collaborative and iterative annotator-involved discussions. The annotators reached good agreement, indicating that the taxonomy and codebook were understandable and useful.

However, to rigorously evaluate the quality of the dataset and the performance of models that build on it, the models should be evaluated in practice with different text types and languages, as well as compared and combined with models trained on different datasets, i.e. Guest et al. (2021). Because online misogyny is a sensitive and precarious subject, we also propose that the performance of automatic detection models should be evaluated with use of qualitative methods (Inie and Derczynski, 2021), bringing humans into the loop. As we found through our continuous discussions, online abuse can present in surprising forms, for instance the denial that misogyny exists. The necessary integration of knowledge and concepts from relevant fields, e.g. social science, into NLP research is only really possible through thorough human participation and discussion.

Acknowledgement

This research was supported by the IT University of Copenhagen, Computer Science for internal funding on Abusive Language Detection; and the Independent Research Fund Denmark under project 9131-00131B, Verif-AI. We thank our annotators Nina Schøler Nørgaard, Tamana Saidi, Jonas Joachim Kofoed, Freja Birk, Cecilia Andersen, Ulrik Dolzyk, Im Sofie Skak and Rania M. Tawfik. We are also grateful for discussions with Debora Nozza, Elisabetta Fersini and Tracie Farrell.

Impact statement: Data anonymization

Usernames and discussion participant/author names are replaced with a token @USER value. Annotators were presented with the text of the post and no author information. Posts that could not be interpreted by annotators because of missing background information were excluded. We only gathered public posts.

Annotators worked in a tool where they could not export or copy data. Annotators are instructed to flag and skip PII-bearing posts.

All further information about dataset creation is included in the main body of the paper above.

References

Hwijeen Ahn, Jimin Sun, Chan Young Park, and Jungyun Seo. 2020. NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-lingual Transfer. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1576–1586.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Amnesty International. 2017. Amnesty reveals alarming impact of online abuse against women. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarmin-impact-of-/online-abuse-against-women/. Accessed: Jan, 2021.

Analyse & Tal. 2021. Angreb i den offentlige debat på Facebook. Technical report, Analyse & Tal.

Astrid Skov Andersen and Maja Langberg. 2021. Nogle personer tror, at de gør verden til et bedre sted ved at sende hadbeskeder, siger ekspert. *TV2 Nyheder*.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, volume 10859, pages 57–64. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Dorthe Bleses, Werner Vach, Malene Slott, Sonja Wehberg, Pia Thomsen, Thomas O Madsen, and Hans Basbøll. 2008. Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, 35(3):619.

Bernadette Campbell, E. Glenn Schellenberg, and Charlene Y. Senn. 1997. Evaluating Measures of Contemporary Sexism. *Psychology of Women Quarterly*, 21(1):89–102. Publisher: SAGE Publications Inc.

Casey Newton. 2020. Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. *The Verge*. Accessed: Jan, 2021.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 13–22, New York, NY, USA. Association for Computing Machinery.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Danske Kvindesamfund. 2020. Sexisme og sexchikane. https://danskkvindesamfund.dk/danskkvindesamfunds-abc/sexisme/. Accessed 2021-01-17.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media*, 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73, New York, NY, USA. Association for Computing Machinery.

Bo Ekehammar, Nazar Akrami, and Tadesse Araya. 2000. Development and validation of Swedish classical and modern sexism scales. *Scandinavian Journal of Psychology*, 41(4):307–314. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9450.00203.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *IberEval@ SEPLN*, pages 214–228.

Mark A Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes and tools. In *Handbook of Linguistic Annotation*, pages 167–191. Springer.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media. *Applied Sciences*, 10(12):4180. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347. Publisher: Association for Computing Machinery (ACM).

Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Peter Glick and Susan Fiske. 1996. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, 70:491–512.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 229–233, Troy New York USA. ACM.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Nanna Inie and Leon Derczynski. 2021. An IDR Framework of Opportunities and Barriers between HCI and NLP. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 101–108.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2018. Cross-Domain Detection of Abusive Language Online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joseph Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wijaya, Xin Wang, Yong Zhang, and Morteza Dehghani. 2018. The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. Technical report, PsyArXiv. Type: article.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2020. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *arXiv:2012.12305 [cs]*. ArXiv: 2012.12305.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The Lacunae of Danish Natural Language Processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carmen Martinez, Consuelo Paterna, Patricia Roux, and Juan Manuel Falomir. 2010. Predicting gender awareness: The relevance of neo-sexism. *Journal of Gender Studies*, 19(1):1–12.

Barbara Masser and Dominic Abrams. 1999. Contemporary Sexism: The Relationships Among Hostility, Benevolence, and Neosexism. *Psychology of Women Quarterly*, 23(3):503–517. Publisher: SAGE Publications Inc.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv:2006.04884 [cs, stat]*. ArXiv: 2006.04884.

Hala Mulki and Bilal Ghanem. 2021. Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative Evaluation of Label-Agnostic Selection Bias in Multilingual Hate Speech Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study. *Information Processing & Management*, 57(6):102360.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label Categorization of Accounts of Sexism using a Neural Framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. "O'Reilly Media, Inc.". Google-Books-ID: A57TS7fs8MUC.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Sima Sharifirad and Alon Jacovi. 2019. Learning and Understanding Different Categories of Sexism Using Convolutional Neural Network's Filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Lea Skewes, Joshua Skewes, and Michelle Ryan. 2019. Attitudes to Sexism and Gender Equity at a Danish University. *Kvinder, Køn & Forskning*, pages 71–85.

Statista. 2020. Denmark: most popular social media sites 2020. Accessed: Jan, 2021.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying Generalisability across Abusive Language Detection Datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Janet Swim, Kathryn Aikin, Wayne Hall, and Barbara Hunter. 1995. Sexism and Racism: Old-Fashioned and Modern Prejudices. *Journal of Personality and Social Psychology*, 68:199–214.

Francine Tougas, Rupert Brown, Ann M. Beaton, and Stéphane Joly. 1995. Neosexism: Plus Ça Change, Plus C'est Pareil. *Personality and Social Psychology Bulletin*, 21(8):842–849. Publisher: SAGE Publications Inc.

Francine Tougas, Rupert Brown, Ann M. Beaton, and Line St-Pierre. 1999. Neosexism among Women: The Role of Personally Experienced Social Mobility Attempts. *Personality and Social Psychology Bulletin*, 25(12):1487–1497. Publisher: SAGE Publications Inc.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300. Publisher: Public Library of Science.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Rex E Wallace. 2005. *An introduction to wall inscriptions from Pompeii and Herculaneum*. Bolchazy-Carducci Publishers.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of Politically Biased Data on Hate Speech Classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). *arXiv:2006.07235 [cs]*. ArXiv: 2006.07235.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis. *arXiv:2005.12423 [physics]*. ArXiv: 2005.12423.

A Appendices

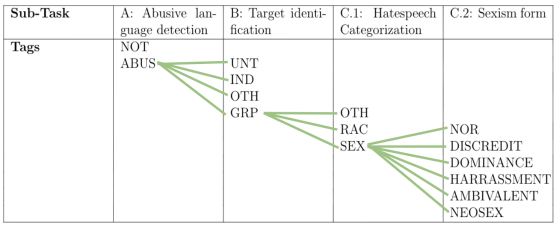
A.1 Annotation Codebook

General Rules for Annotators

- The focus of the annotation task is on the **whole post**. Some words and hashtags depend on their contextual use, if they are meant offensive, not.
  For example:
  (1) "hykleriske", "ad helvede til","føj", "sgu", "pisse", "fanden", "eddermame", "Hold kæft", "liderkarl", "bolle", "åndssvag" eller "løgner"
  In case these words appear without any context in a post, i.e. "Hold nu kæft", the post is not abusive.
  Quotes: Quotes are considered as always context-dependent. The author uses someone else words and agree with them, not. Bordercase: If a quote and only if is used without any further comment, between two cases are distinguished:
  1) Quote contains profanity: Labeled as: ABUS/UNT, i.e.
  (2) Copypasta textual memes: i.a. Navy Seal Copypasta
  2) Quote contains vague abuse without any profanity/slurs: not ABUS, intention of the author is unclear why the quote is posted.

- **No observations on top of the post**, just the text of the post is relevant for the evaluation.
  Examples for being not abusive just by the post itself:
  (3) "It is best that they stay there and not come back."
  (4) "Jo og hendes der gambler med Danskernes penge"
  (5) "hvorfor ikke sætte navn på manden ??"

Annotation scheme



Precedence of labels: For each post a label is chosen in Sub-Task A according to the annotation scheme. Depending on the chosen label, further labels (Sub-Tasks B and C) may need to be selected following the hierarchically annotation scheme above (green lines). The determining label addressed by the post should be selected.

For example, the primary abuse of this posts adresses racism, where the chat participant is offended by the fact of being from a "dansk/afghansk kultur":

(6) "@USER Du ser sexistiske spøgelser alle vegne, fordi du kommer fra en dansk/afghansk kultur, hvor overgreb mod kvinder er almindeligt accepteret og derfor en del af selvforståelsen."

SubTasks and Tags

SubTask A: Abusive language detection

Generally, posts containing abusive language include insults, threats, any type of untargeted profanity. (**ABUS/NOT**) Specifically, a post is abusive if it:

- uses slurs, clear abusive expressions (In case of censorship, i.e. "p*s","fu..", the actual slur has to be clear).

(7) "kælling", "lort", "klamme svin", "sindssyge", "idiot", "fucked/fucking", "wtf/what the fuck", "luder"

- attacks a person, minority to cause harm, repetitiveness,, an imbalance of power (examples see subTask B).

- promotes, but does not directly use abuse language, violent crime, i.e. agreeing with a abusive quote by "#præcis".

(8) "@USER: hørt på tribunen: jeg elsker alle dansker men pigerne har en klam personlihed. ludere #præcis."

- contains offensive criticism without a well founded argument/ backed-up fact.

(9) "Ja. Pippi Langstrømpe fx. Mega negativt portræt og meget undertrykt af Patriarkatet."
not offensive criticism - NOT abusive:
(10) "det ville være dejligt hvis tvangsfjernelser omskæring gensidig forsørgelse barnebrude syge borgere fik lige så meget hjælp af offentligheden og medierne."

- blatantly misrepresents truth, seeks to distort views on a person, minority with unfounded arguments/ claims.

(11) "Mænd kan ikke blive medlem hos de radikale, det ender med noget Rod!!!"

(12) "Feminist partiet, vil blot have fjernet mænd fra parti toppen, bruger derfor beskidte kneb"

(13) "Lad nu Morten være i fred. Men et par sygemeldinger fra Sofie Carsten Nielsen og Lotte Rod mangler vi. For Sofie og Lotte har sider på spring for at finde noget på Morten, bare for at få Formands posten. De skulle skamme sig. Må da ikke håbe folk stemmer på dem/ hende til næste valg."

- shows support of problematic hash tags, screennames.

(14) "#KriminelleUdlændinge".

Explanations various hashtags:

- "#sikkerhedssele": Disadvantage of women in rights/ opportunities because of male-dominated development&research

- "#såfuckdig": used to express that women victimize themselves

- "#mændsligestilling": used against feminism but also just showing men inequality in some cases

- negatively, positively stereotypes in an offensive way (examples, see subtask C: Stereotype & Objectification).

- defends xenophobia, sexism (examples, see subtask C).

- seeks to silence a person, minority by consciously intending to limit the person's freedom of speech.

(15) "Stop nu alt det fnidder !!"

- is ambiguous (sarcastic/ ironic), and the post is on a topic that satisfies any of the above criteria (see various examples in the Appendix bordercases).

SubTask B: Target identification

An abusive post can classified as untargeted (**UNT**), targeted (**IND/GRP/OTH**). Untargeted posts (**UNT**) contain nontargeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language.

(16) "REMOULADE, KÆLLINGER!"

(17) "Indtil lortet begynder at hvæse og går amok..."

(18) "Er de af glas så smid dem i flaskekontaneren du vil vel ikke have andre skal bruge det lort du vil jo ikke selv."

Targeted posts can be towards a specific individual person/ persons that is/ are part of the chat, a person the conversation is about (**IND**). The person can be named (i.e. @USER), unnamed.

Example for a specific person, unnamed:

(19) "største luder på dansk jord."

Example for a specific person, named:

(20) "@USER , det er ikke dig der dømmer. Du lovgiver. Hvis du mener noget så må du gå i Folketinget. Føj for satan, du er mega klam."

Example for a specific person the chat is about:

(21) "@USER Prøv at se hende i øjnene. Hun kan lyve overfor den danske befolkning uden, at blinke med øjnene... imponerende. Men kikker du RIGTIGT efter, så vil du se satan selv, i hendes øjne! Ondskabsfulde kælling!"

(22) "Havde Gru eksisteret i dag, var Joe Biden sikkert på forsiden med en lille pige."

Posts can also contain any language that disparages a group of people united by race, color, ethnicity, gender, sexual orientation, nationality, religion,, other social identity (**GRP**). This can target the group in general, an individual group members receiving abuse because of their group membership.

(23) "Det er sgu flot så kan de komme og sprede smitten rundt gud fader bevar os."

If a post is still targeted but does not belong to one of the other categories, i.e. targeting an organisation, country (if it is not used to refer to the nationality of persons), party, it should be labeled as **OTH**.

(24) "Jeg er sgu lidt forvirret... hvad var det Danske Bank lavede lort i, denne her gang?"

This also includes undefined groups outside the chat, i.e.

(25) "Kære alle idioter. Samtykke og #MeToo er også frigørende for mænd."

Additional clarifications about certain groups:

(26) "medier" = company

(27) "immigrants" = group based on race

(28) "muslimer" = racism

SubTask C: Hate Speech Categorization

A post containing abusive language against a group can be sexist (**SEX**), racist (**RAC**), of other nature (**OTH**). Sexist content can be described as prejudice, stereotyping,, discrimination typically

against women, while racism is on the base of racial, ethnic characters of a group.
Others can be: LGTB, sexism against men

Racist posts can be:
(29) "@USER Hvad med et forbudt kun for muslimer.!:) Dette vil være en løsning."
(30) "@USER Migranter, der ikke har fået asyl eller flygtningestatus kan smides ud uden problemer. det er lykkesøgere så man kan følge dem til den tyske grænse,og de kan så forsøge der ellers må de rejse videre til næste land, og prøve lykken DER ER INGEN KRAV OM, AT DE SKAL FORSØRGES AF DK."

Sexist posts are categorized in:

- **Stereotype & Objectification** (**NOR**), if a normative held but fixed and oversimplified image, idea (Stereotype), description of person's physical appeal and/or comparisons to narrow standards (Objectification) is used in an offensive way.

(31) "Kvindefoodbold , ??? Hummmmm, gabbbbbb"
(32) "Hysterike kællinger"
(33) "Du høre til i køkkenet eller i vaskekælderen."

- **Benevolent** (**AMBIVALENT**), if the author uses a strong subjective positive sentiment/ stereotypical picture with men, women often disguised as a compliment (benevolent sexism), i.e.
  "They're probably surprised at how smart you are, for a girl", there is a reverence for the stereotypical role of women as mothers, daughters and wives: 'No man succeeds without a good woman besides him. Wife, mother.'
  Examples are:

(34) "Spørg Tina Bilsbo. Hun havde nosser og tog konsekvensen, da hun ikke fik det samme som de mandlige."
(35) "Tag det som end mand."
(36) "Piger og kvinder er de meste udsatte borger både i Danmark og på det globale plan.! Så din offer rolle virker ikke."

- **Dominance** (**DOMINANCE**), if the author assert the superiority of men over women to highlight gender inequality without using a stereotypical picture.

(37) "Kvinder er mindre intelligente end mænd."
Accusing somebody as "lille pige" is categorized under dominance (not stereotyped) of the male gender over the female:
(38) "@USER Det var den gamle. Jeg har udviklet mig. Du flæber stadig som en lille pige."

- **Discredit** (**DISCREDIT**), if the author disgrace, humiliate woman with no larger intention.

(39) "Kvinder er fucking korrupte, sådan er det desværre bare."
(40) "Danes for Trump er ikke kæmpe fan af bingoværten fra Jylland, men er nu alligevel lidt triste over den måde, som en flok hysteriske kælliger har fået manden slået tilbage til bingostudiet på."

- **Sexual Harassment & Threats of Violence** (**HARRASSMENT**), if actions as sexual advances, requests for sexual favours, harassment of sexual nature is used; intent to physically assert power over women through threats of violence.

(41) "Hvorfor er særligt kvinder tolerant ift. islam ? Mange kvinder må tilsyneladende have drømme om at blive behandlet som slave eller luder."

- **Neosexism** (**NEOSEX**): Egalitarian values about women and the justification of woman abuse. Authors:
  1. deny the existence of discrimination against women.
  2. resent complaints about discrimination.
  3. resent "special" favours for women.

  In general, neosexism measures more attitudes toward feminist changes in society not direct attitudes towards women. It follows the idea of: "women are no longer discriminated in our society." People expressing neosexist beliefs have an interest in opposing to social policy changes that would benefit women and keeping the status quo although they may maintain non-traditional gender roles.

Example for (1.) questioning the existence of discrimination:

(42) "Kan I pege på forskning der viser at barslen er grunden til at mødrene går glip af forfremmelser?."

Example for (2.) resent complaints about discrimination:

(43) "Jeg føler ofte folk har behandlet mig bedre og talt pænere til mig
fordi jeg var en pige, så jeg har ret svært ved at tage det seriøst når folk mener at kvinder er sååå diskriminerede imod i den vestlige verden."

Including authors demonstrating that "men are victims of the feminism movement":

(44) "Der er nu mange middelaldrende mænd, som er endt i en prækær situation som 'den pressede mand' tæt på bunden. Husk at skrive om mænd der ikke er i medieeliten."

(45) "Klassisk. Hvis det er en ulempe for kvinder er det samfundets skyld. Hvis mænd, så må det jo være deres egen. Sexisme trives godt på den feministiske fløj."

But barely demonstrating men inequality is NOT neosexism. It does not deny the existence of discrimination of women, i.e.

(46) "Hvad med alle de som er soldat og er faldet i kamp? Der mange flere mænd som er død i kamp! Hvorfor hylder man ikke dem enkeltvis? Der fandme intet ligestilling der. . . "

Example for (3.) resent "special" favours for women:

(47) "Man kan ALTID finde en ting at pege på, uanset kontekst, hvor kvinder er dårligere stillet. FX whatabout: Smerter! Ingen andre steder at konkludere sig hen end partriarkat og systematisk kvindeundertrykkelse."

*(48) "Det er også kendt at det først er indenfor de seneste få år at kvinder er blevet nervøse for at være alene med fremmede mænd langt fra andre mennesker... *(face with rolling eyes)* Aldrig været et issue før i historien."*

A.2 Danish Misogyny (empirical)

- The oldest women's organization in Denmark Danske Kvindesamfund (2020) defines sexism against women as *"Sexisme er en fordom eller diskrimination på baggrund af køn, især i forhold til nedvurdering af kvinder."* (Sexism is a prejudice, discrimination based on gender, especially in relation to the downgrading of women.)
- Skewes et al. (2019) present a survey at a Danish university and their findings exhibit a correlation of modern sexism scale and the attitude "enough, too much was being done for gender equity".
- Ekehammar et al. (2000) proves the existence of modern sexism attitudes in Sweden with stronger means for men.
- The modern sexism scale is based on the modern sexism theories studied in North America and their application to the European context already proven by Masser and Abrams (1999) (study undertaken in Great Britain).

A.3 Search Keywords

- ambigious keywords: voldtægt, synd, helvede, lækker, dødt, sæk
- slurs from Sigurbergsson and Derczynski (2020): fisefornem, hjemmefødning, kvindeagtig, ludder, papmor, pigebarn, pigefnidder, plasticmor, tyskertøs, pattebørn, kvindermenneske, svabrefjams
- from articles/interviews: luder, møgsæk, gruppevoltægt, kælling, lille pige, dumt svin, klam
- translated from previous work: "så god som en mand", "som en mand", "til en pige", "smart til en pige", "kærlighed til en kvinder", "intelligent til en pige", #adaywithoutwomen, "en dag uden kvinder", "#womensday", "#everydaysexism", "#weareequal"
- by pattern recognition from posts: "#MeToo", "#getbackinthekitchen", "som end mand", "gør noget rent", "jeg er jo en mand"', '"kvinder er", "til en pige", "en mand som", "lille pige", "dumt svin", "høre til i køkkenet", "vi kvinder", "men kvinder", "mænd der siger"', "#Mændsligestilling"
- for related topic-search: 'kvinder', 'sexisme', 'voldtaget', 'sex', 'skræmmende','mediechefer','trussel','indvilligede', 'mandlige kolleger','sexisitisk','mediebranchen', 'sexbeskeder','kvindelige'