


ARTICLE

# A study towards contextual understanding of toxicity in online conversations

Pranava Madhyastha<sup>1,2</sup> , Antigoni Founta<sup>2</sup> and Lucia Specia<sup>2</sup>

<sup>1</sup>Department of Computer Science, City University of London, London, UK and <sup>2</sup>Department of Computing, Imperial College London, London, UK

**Corresponding author:** P. Madhyastha; Email: [pranava.madhyastha@city.ac.uk](mailto:pranava.madhyastha@city.ac.uk)

(Received 31 August 2021; revised 20 June 2022; accepted 21 June 2022)

## Abstract

Identifying and annotating toxic online content on social media platforms is an extremely challenging problem. Work that studies toxicity in online content has predominantly focused on comments as independent entities. However, comments on social media are inherently conversational, and therefore, understanding and judging the comments fundamentally requires access to the context in which they are made. We introduce a study and resulting annotated dataset where we devise a number of controlled experiments on the importance of context and other observable confounders – namely gender, age and political orientation – towards the perception of toxicity in online content. Our analysis clearly shows the significance of context and the effect of observable confounders on annotations. Namely, we observe that the ratio of toxic to non-toxic judgements can be very different for each control group, and a higher proportion of samples are judged toxic in the presence of contextual information.

**Keywords:** Natural language in multimodal and multimedia systems; Corpus annotation; Understanding toxic language

## 1. Introduction

Social media websites provide an important platform for conversational engagement. However, these platforms can suffer from a variety of toxic conversational behaviours, including abusive, hateful, antisocial attitudes, which result in adverse effects ranging from conversation derailment and disengagement to life-threatening consequences. Research in understanding and forecasting such behaviours has received significant attention in recent years (Bamman, O'Connor, and Smith 2012; Barker and Jurasz 2019; Gomez *et al.* 2020). This has also spurred a sustained discussion in policy and regulatory forums across various governmental bodies, including the UK's Online Harms white paper<sup>a</sup> and the congressional hearing on content moderation in the US.<sup>b</sup> Most research has focused on overt forms of toxicity, where the content is usually analysed independently of the conversational context (Vidgen *et al.* 2019). Additionally, in current research, the perception of toxicity in content and/or its labelling have been constrained to either a general pool of crowd-workers or an abstract view from experts, without taking into account important factors that can affect perception, including various demographic features, such as age or gender, and political views (Waseem 2016; Davidson *et al.* 2017; Kiela *et al.* 2020). The variety of confounders can potentially result in biased annotation (Davidson, Bhattacharya, and Weber 2019; Sap *et al.*

<sup>a</sup><https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper>

<sup>b</sup><https://www.c-span.org/classroom/document/?17312>

2019a). The bulk of work that focuses on understanding toxic online content has predominantly cast this as the task of predicting toxicity. However, due to above-mentioned biases in annotations, the predictive models are prone to amplify the biases. Finally, there is much less understanding of ‘what in the content’ leads to the decision of the sample being annotated or predicted as toxic.

Our primary goal in this paper is to study toxicity in comments made by users of online platforms on a piece of content in a controlled way, and thus shed some light on the above issues. Our main hypothesis is that understanding of toxicity in online conversations can only be done in context. Here, we specifically define context in terms of (i) additional content related to the item we aim to analyse, be it textual or in other modalities; and (ii) external factors that influence perception of toxicity under confounding factor, such as gender, age and political orientation.

Towards this study, we introduce a dataset<sup>c</sup> of online content in English, which contains comments with textual and multimodal context. The dataset is comprised of 1500 distinct samples, with multiple annotators per sample, two main sample-level setups – comment-only and comment-plus-context – and three control groups: male vs female, right-leaning, left-leaning versus centre-leaning, and under 30 versus 30 and over years of age. These control groups by gender, political orientation and age are aimed to quantify the extent to which the perception of toxicity varies across different annotators. In addition to annotations for toxicity, we also collect annotations for sarcasm, as this is known to affect toxicity judgements. Finally, the annotation also contains annotator rationales, that is reasons (within the comment or the context) for the decision of toxicity. In total, the dataset has 21,000 samples labelled for toxicity and sarcasm with rationales. Our main contributions in the paper can be summarised as follows:

- We present the first annotated dataset in English where we control for a variety of factors, including the context that leads to comments, gender, age and political orientation (Section 4).
- We study the extent to which observable confounders modulate the annotations (Section 5).
- We study the extent to which context influences the annotation process (Section 6).

Our data annotation and analysis clearly demonstrate that the perception of toxicity is contextual.<sup>d</sup> For example, the presence of context increases the proportion of samples judged as toxic. Further, the annotations can significantly vary depending on the different control groups. For example, content annotated by people who identify themselves as right-leaning is considered significantly less toxic.

## 2. Background and related work

### 2.1 Toxicity in online conversations

Characterising toxic content is a difficult challenge as toxicity is a broad concept that can include threats, racial slurs, extremist views, sexism, insults, hateful, and derogatory comments, which can be both directed at specific users or at communities. In this context, Jurgens, Chandrasekharan and Hemphill (2019) categorise toxic online content into two categories – (a) Overt, physically dangerous content: content that may or may not use abusive or hateful language and may include personally identifiable information to cause harm (e. g., doxxing). Content that promotes behaviours of criminal harassment, human trafficking, paedophilia or explicitly incites mobs for violence – all these constitute the category of overt and physically dangerous content; (b) Subtly harmful content: content that is linguistically subtle, where harm is implicit. We note that the latter category has been predominantly disregarded and is among the difficult sets of samples for annotation. Our work focuses mostly on the latter category.

<sup>c</sup>The dataset will be released on publication of the paper with CC-BY-ND-NC license

<sup>d</sup>the data are made available here: <https://osf.io/9p6cz/>

There has been an increasing amount of work towards studying toxic online conversation in the last few years, which has focused on overt and physically harmful online behaviours (Waseem 2016; Vidgen and Derczynski 2020). The topic has been studied under various lenses, including that of ‘abusive language’ (Nobata *et al.* 2016), ‘harmful comments’ (Faris *et al.* 2016), ‘toxic language’ (Jacobs *et al.* 2020), ‘online hate speech’ (Davidson *et al.* 2017) and incivility (Kumar *et al.* 2018). Further, recent work has also analysed online conversation in the context of cyberbullying and trolling (Yin *et al.* 2009; Cheng *et al.* 2017; Kumar, Cheng, and Leskovec 2017; Liu *et al.* 2018). Significant efforts have been made towards conducting shared tasks to increase community participation and benchmarking towards the task of detecting toxic behaviours. These include Zampieri *et al.* (2019, 2020) and Basile *et al.* (2019) among others. There are various taxonomies for both overt and covert types of toxic conversations. We refer the reader for a detailed analysis of taxonomies to Vidgen *et al.* (2019).

Previous work has studied different aspects of toxic online conversation, including the linguistic variation, such as understanding linguistic changes in user behaviour in the context of community norms (Danescu-Niculescu-Mizil *et al.* 2013) and understanding the influence of community norms in the context of community moderation (Chandrasekharan *et al.* 2018). Their study indicates the presence of macro-, meso- and micro-norms that are specific and unique to particular subreddit communities and modulate different types of disparate conversation. Norms play an important role in moderation of online communities. In our work, we are interested to explore how would a general populace perceive such moderated content. This helps identify potentially outward-facing moderation strategies.

## 2.2 Datasets

A variety of online content datasets have been created for studying toxicity in online conversations. However, the samples in most existing datasets are predominantly collected with the purpose of building and evaluating supervised machine learning models, to predict toxicity, rather than to obtain a deeper understanding of the topic. Most of the datasets have samples that consist of predominantly textual comments and with sample sizes ranging between 500-500,000.

We briefly describe some of the most popular datasets that have been used for studying toxic language and building predictors. We restrict ourselves to English language datasets and refer the reader to Madukwe, Gao and Xue (2020), Poletto *et al.* (2020), Vidgen and Derczynski (2020) for a more comprehensive review of these and other collections. Using Yahoo News Groups, Warner and Hirschberg (2012) present a dataset based on comments and hyperlinks from the American Jewish Society. The comments are manually annotated using seven categories: anti-semitic, anti-black, anti-Asian, anti-women, anti-Muslim, anti-immigrant and other forms of hateful behaviour (anti-gay, anti-white etc.). Djuric *et al.* (2015) use comments from Yahoo Finance which are annotated by crowd-workers for the presence or absence of hateful comments, while Nobata *et al.* (2016) use comments from Yahoo Finance and Yahoo News annotated for abuse, simply whether they are abusive or clean, through crowdsourcing.

Using Twitter as platform, Burnap and Williams (2016) investigate aspects of cyber-hate, with an emphasis on a set of triggers including sexual orientation, race, religion and disability on tweets. The dataset was annotated by crowd-workers who were answering if the sample is offensive or antagonistic in terms of the aforementioned triggers. Waseem and Hovy (2016) introduce a dataset with tweets annotated by the authors and validated by experts, where the focus was on annotating tweets that showcase racism or sexism. A further extension enhanced the dataset with annotations from both the experts and crowd-workers and analysed the influence of expertise in the annotation process (Waseem 2016). Founta *et al.* (2018) present a large dataset of 100,000 tweets, focused on annotations of abusive, hateful and spam-based categorisation by crowd-workers. The three categories are the result of an iterative preliminary analysis, where the

authors experiment with a wide variety of often interchangeable labels, to distinguish the relationship between them and select the most representative. The various labels are either eliminated or merged, through statistical analysis, to produce a smaller robust set which covers a range of abusive behaviours.

In terms of datasets including modalities other than text, Gomez *et al.* (2020) use Twitter to curate a multimodal – image and text-based – meme dataset, where memes are annotated by crowd-workers on six categories: racism, sexism, homophobia, religion-based hatred and attacks on other communities. A similar work by Kiela *et al.* (2020) presents a synthetic meme dataset, where the focus is on categorising samples as multimodally hateful, in which case the image and the associated text are hateful, or unimodally hateful, in which case either the image or the text is hateful.

A few recent works have focused on studying toxicity in a contextual setting, we here list the most prominent works. One of the earliest works to study the importance of context is by Gao and Huang (2017), where the authors present an annotated corpus constructed using articles and the corresponding discussion on a news website. This work considers the preceding discourse and the corresponding news item as relevant context. In Qian *et al.* (2019), the authors collect a dataset of hateful comments and crowdsource interventional responses to these comments, from a small set of subreddits. The annotation task aims to identify toxic comments, as well as create human-written responses to the total conversation, which would intervene to the discussion and hold back hate. The annotations are collected through crowdsourcing, and the contextual intent is also provided to the annotator, in matters of title and content from the original submission. Voigt *et al.* (2018) release a large-scale raw dataset (not annotated) with Reddit comments and the associated user's gender. Pavlopoulos *et al.* (2020) also create a context-aware toxicity dataset from conversational comments from Wikipedia talk pages discussion pages. The authors are interested in similar questions as to whether context can modulate human judgements of toxicity and the utility of context on machine learning models. However, the contextual information is local and does not take into account global subject content under discussion. Shen and Rose (2021) investigate the influence of annotators' political beliefs and familiarity of the source towards annotations of identifying political orientation of Reddit posts. Chung *et al.* (2019) present a dataset that collects both hateful and toxic messages along with potential repairs that can provide counter-narrative information with fact-based information and non-offensive language to de-escalate hateful discourse. Ljubešić, Fišer and Erjavec (2019) present a dataset in both Slovenian and English annotated for a variety of annotations for socially unacceptable discourse. The dataset consists of conversational content from Facebook platform, such that the entire discourse thread is presented to the annotator during annotation for each potentially toxic comment.

We note that the predominant approach in curating samples from social media platforms is by using content that contains words from a small seed lexicon of hateful words. Most datasets focus on annotating comments, where the annotators are not given any context or additional information, such as what the comment refers to, or any history of conversation. The comments are usually taken as an independent entity. Even in the two multimodal datasets, the sample is a unique piece of content which, in this case, happens to have both image and text in it. Our dataset is based on comments from Reddit, where we focus on understanding the perception of toxicity. In contrast to other datasets, our work is a controlled study on how the labels change in the presence and absence of different types of context. We are also interested in understanding the influence of demographic information on annotations.

### 3. Data annotation strategy

#### 3.1 Definition of toxicity

Previous work in this domain has focused on fine-grained distinctions of toxic content in the form of hate speech, abusive language and other profanity-related terms. We note, however, that the

definitions used for the fine-grained aspects of toxicity have been fairly ambiguous and have been used interchangeably (Founta *et al.* 2018; Vidgen *et al.* 2019). In addition, recent work has studied the prevalence of biases due to annotation artefacts in datasets, including those on toxic online content (Davidson *et al.* 2019; Sap *et al.* 2019b). Among the reasons for the presence of such biases, one prominent aspect is the complex and stratified definitions of fine-grained forms of toxic online content. It has been shown that complex definitions need experts or trained annotators and may yield biased annotations, when seeking annotations from crowd-workers (Waseem 2016).

In this paper, we are interested in a simpler and broader definition of toxic content, as we focus on capturing the varying perceptions of toxicity. To this end, we conducted various pilot experiments on the crowdsourcing platform with a few samples to study annotators' responses while varying the toxicity definition. Our goal was to estimate the difficulty of content annotation, based on a variety of definitions. After five rounds of experiments, we observed that a broader and simpler definition was conducive to annotations and also received positive feedback from the crowd-workers. We thus settled on the definition of toxicity proposed in Wulczyn, Thain and Dixon (2017) and Davidson *et al.* (2017): "any form of content that is hurtful, derogatory or obscene, such that it can hurt a person either physically or emotionally, for example content threatening a person or spreading hateful messages against a person or a group".

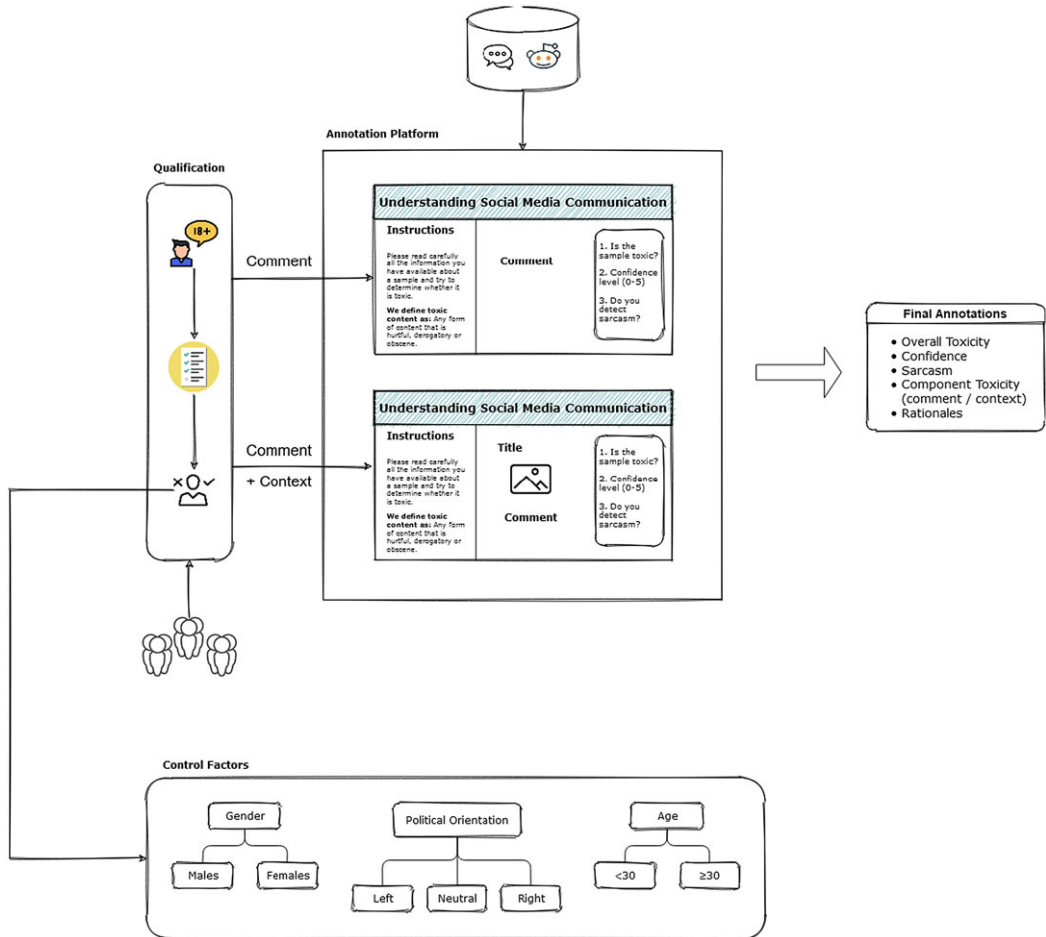
### 3.2 Contextual multimodal content

A fundamental concept of online social media platforms is the interaction among users and their use as conversational platforms, typically through comments or replies to original submissions. Therefore, the *contextual* aspect of these interactions should be considered in social scientific research, especially given the nature of toxicity. This has also been emphasised in previous work, such as Vidgen *et al.* (2019). In the previous section, we define toxicity as content that can hurt a person or a group, hence we presume that there is usually a target.

Despite the prominent role of context, it is not common in previous work to consider contextual information when annotating content. This is one of the most important contributions of our work – studying the effect of context on toxicity annotation while also focusing on considering demographic factors influencing annotations.

Furthermore, social media content exists in multiple modalities including textual, visual (such as images or videos) and audio forms. Various platforms support different modalities, and some modalities can be more prominent in conversations depending on the platform, such that the textual modality may not necessarily play the most important role. Both Kiela *et al.* (2020) and Gomez *et al.* (2020) present multimodal datasets that contain memes or pictorials with textual information as samples. However, we classify these as a different type of content, which has a very different communication intent.

To measure the effect of context, we run two parallel experiments where the only difference is the availability of context during annotation for one of the experiments. In the first experiment, we provide comments obtained from Reddit to a group of annotators and ask the participants to annotate whether they believe the comment (which is only comprised of text) is toxic under our definition. For the second experiment, we provide the same samples and annotation setup, only this time participants have access to the additional context from the submission the comments belong to. This context includes the *title* (in the form of text), along with the *image* of the original submission, and the annotators are asked to decide upon the toxicity of the sample under the same definition, but now given the context. We restrict our definition of conversational context to this content since it is what prompts users to post comments on the platform. The context thus acts as the communication intent and we hypothesise that this adds important information to decide on the toxicity of the sample; however, this may further increase the subjectivity in analysing the samples due to the additional source of information. We present an overview of our experimental setup in Fig. 1 and instances of the samples in Fig. 2.



**Figure 1.** Overview of the experimental setup. In order to participate in the experiments, workers first need to pass a series of qualifications. Once they successfully qualify, they are separated into control groups based on the confounding variables of each experiment. Participants are then given access to the annotation platform and, based on the experiment type and provided information, are asked a series of questions related to toxicity. The final annotation scheme is rich in information, approaching toxicity in a holistic manner.

In the rest of the paper, we consider the two different set of experiments as two different projects and represent them with the letters C and CTM, where C stands for Comment-only tasks, that is no context was presented to the annotators, and CTM stands for Comment-Title-Media, that is samples that are enriched with contextual information.

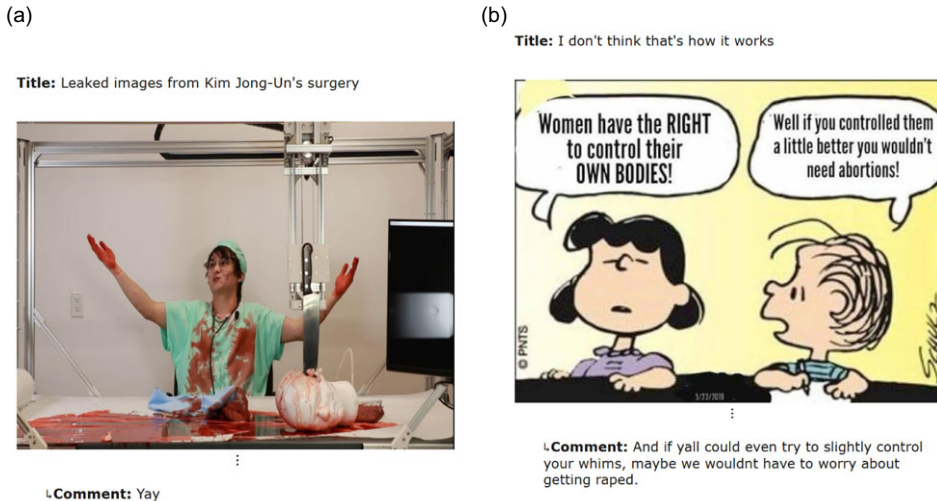
### 3.3 Sample selection

Our dataset is based on conversations from the Reddit platform. Most existing works use a form of seed lexicon or a set of seed topics to build the dataset. In this study, we base our collection on the strategy from Bamman *et al.* (2012).

We gathered a list of 400 most popular subreddits based on the comment traffic on the subreddits.<sup>e</sup> Like Twitter and other social media services, Reddit provides developers with open APIs

<sup>e</sup>As listed on <https://pushshift.io/> on September 30, 2019,





A sample that is highly toxic, yet toxicity originates from context rather than the comment on its own. A toxic sample where the comment seems foul, however it becomes clearer in the presence of context.

Figure 2. Examples from the dataset.

to build services, including access methods to storing a snapshot of a post and the corresponding comments. We also exploit Pushshift API (Baumgartner *et al.* 2020) which additionally provides access to retrieve historical changes in Reddit data. This functionality allows to especially track comments and posts that were made on the platform which were then subsequently removed by the subreddit moderator. We note that the associated norms may not always be relating to the ‘toxic’ content, for example a subreddit dedicated for discussion of football-related news may consider comments and posts that discuss political issues to be breaking the norms of that specific subreddit.

In order to build our dataset, we queried Reddit at fixed intervals using both Reddit and Pushshift API to retrieve a sample of messages. Every message in our corpus was initially written between 1st October 2019 and 30 October 2020. We then use Pushshift API to identify messages that have been deleted by the moderator. The API returns the message “[removed]” for content that is removed by the moderator of the subreddit, while it returns the message “[deleted]” for content that is deleted by the user themselves. The comment is usually removed by the moderator as it breaks one or more of the norms associated with the subreddit. Of these, we focus only on content of the type where we have access to a fully formed title, an associated image and the corresponding comment that is removed. In this version of our dataset, we focus on top-level comments. In this study, we deliberately choose top-level comments in order to remove possible confounding due to the discourse. Therefore, every sample in our dataset consists of a title, an associated image and a removed comment.

We observed that there are a substantial set of the samples that were not appropriate for the purposes of annotation. These were messages that were strictly not textual (such as hyperlinks), or the posts were typically inappropriate (such as constituting gore and other bad content). Over the course of the whole year, we sampled over 60,566 piece of toxic content. From this, we randomly sampled, without replacement, a subsample of over 2500 samples. Of these samples, the exclusion criteria for removal were as follows:

- Content devoid of comments that were not predominantly text-based (i.e., comments that mostly comprised of hyperlinks/a single token)

- Samples that broke our institution's ethics guidelines: these were specifically related to imagery that consisted of visually harmful content, such as content showcasing extreme violence or animal abuse or sexually explicit forms of content.
- We also removed content where either the title or the comments were not in English.

We then carefully curated a set of 1500 samples that were deemed appropriate for the crowdsourcing platform.<sup>f</sup> We also note here that the set of subreddits considered consist of subreddits that have left-leaning, right-leaning and politically neutral subreddits.

### 3.4 Annotation setup

We use Amazon Mechanical Turk (AMT)<sup>g</sup> for collecting annotations. AMT is a crowdsourcing website where individuals (or requesters) can collect annotations using a distributed set of crowdworkers. It has been broadly favoured by the research community for data annotation tasks as it has a sufficiently expressive interface for controlled and randomised experiments. Annotations can be obtained relatively in a short period of time due to its large workforce.

AMT offers a set of tools for a multitude of projects. In our work, we exploit some of these existing API tools to implement the final custom platform, which can be seen in Figs. 3(a)–4(b). Our task consists of three parts: the instructions, the practice trials, and the main annotation task.

The instructional part contains information which details the task requirements. Here, we briefly describe the task on a main page, but also offer more details as well as a video and a set of example annotations.

The other two parts of the platform are task-dependent: comment-only (C) versus context-rich (CTM) tasks. For the samples that relate to the comment-only tasks (C) each annotation sample consists only of the textual comment and no other information (Fig. 3(a)). On the other hand, the context-rich tasks include annotation samples with the title of a post, an image originally associated in the post and the comment related to this post (Fig. 4(a)).

The annotation scheme also changes for the two cases. Specifically, the rudimentary annotation framework remains identical in both cases, that is identifying whether the sample is toxic. The details, however, change for the latter, where annotators are asked to judge not only the comment but *also* the context for their toxicity. In both cases, the first annotation task is a simple classification of whether the sample, as a whole, can be considered toxic, that is based on the specific definition of toxicity (described in Section 3.1) workers are asked to judge if the whole sample is toxic. If for any reason, they are unable to decide or have other issues with the sample, they have the option to skip. Even though workers rarely opted to skip in our pilot experiments, we received feedback from the annotators expressing the importance of this option. A participant, for example, mentioned “*Having an option to skip is a very important button I’m glad you included, because for some comments I feel like I don’t have enough authority to really say whether or not something is toxic*”. [sic].

We first present the common setup across both C and CTM experiments. Following the toxicity judgement, we are interested in understanding the degree to which annotators are confident in their decision, so we ask them to provide an estimate from a scale of 0 to 5. We also ask annotators to highlight the portion of the comment text that makes the sample to be perceived as toxic. This allows us to study reasons (i.e., rationales) that led the annotator to identifying the sample as toxic.

During our data curation, we observed that a significant number of comments are sarcastic. As the samples in our dataset are mostly of subtly harmful content, the use of sarcasm to mock or convey contempt is a common observation. Previous work has also suggested that such behaviour is more frequent in subtle and covert forms of abusive and hateful language (Malmasi and Zampieri

<sup>f</sup>this was explicitly done by the first authors of the paper

<sup>g</sup><https://www.mturk.com/>



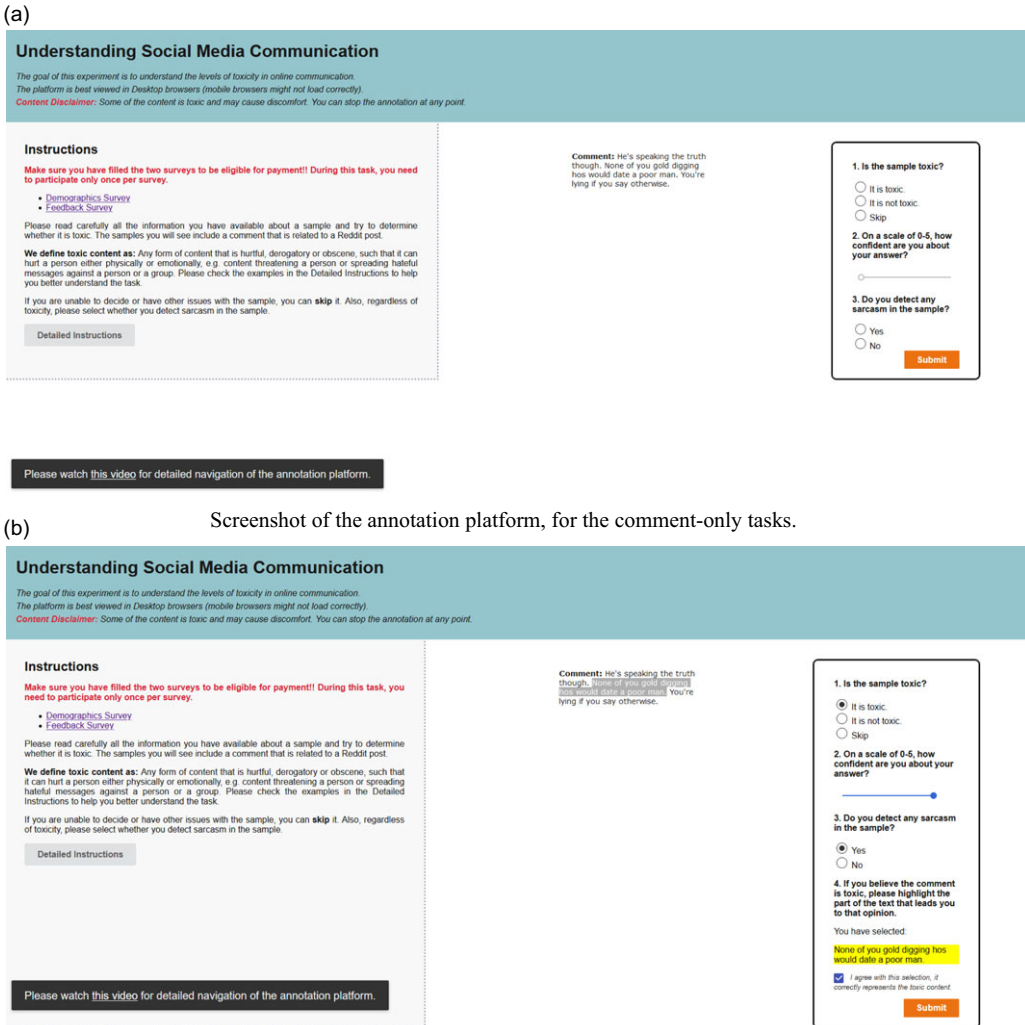


Figure 3. Our annotation platform for the comments-only experiment.

2018). Further, it has also been pointed out that it is difficult to get correct judgements in such cases as it requires a broader knowledge and understanding of the specific communities and commonsense assumptions (Nobata *et al.* 2016). We refer the reader to Vidgen *et al.* (2019) for a more comprehensive survey on the challenges of detecting and annotating subtly harmful content. To allow for such analysis, we further ask the annotators to decide whether the sample was seen as sarcastic.

In our setup, the above format remains consistent across experiments. In addition to the previous questions, for the CTM experiments we further ask the annotators whether the comment and/or context are toxic. Note that, in this case either any of the comment or the context or both can potentially be annotated as toxic. Finally, when the content (in the form of comments) or the context (in the form of titles and images) are considered toxic, we want to further identify the parts of the sample that make it toxic. We also ask the participants to highlight the parts of text

(a)

### Understanding Social Media Communication

The goal of this experiment is to understand the levels of toxicity in online communication. The platform is best viewed in Desktop browsers (mobile browsers might not load correctly).  
**Content Disclaimer:** Some of the content is toxic and may cause discomfort. You can stop the annotation at any point.

**Instructions**

**Make sure you have filled the two surveys to be eligible for payment!! During this task, you need to participate only once per survey.**

- Demographics Survey
- Feedback Survey

Please read carefully all the information you have available about a sample and try to determine whether it is toxic.


The samples you will see include the title of a post, an image embodied in the post, and a comment that is related to this post. Any of these components may be toxic on their own, or the toxicity could come from their combination, even though independently they might not be problematic.

**We define toxic content as:** Any form of content that is hurtful, derogatory or obscene, such that it can hurt a person either physically or emotionally, e.g. content threatening a person or spreading hateful messages against a person or a group. Please check the examples in the Detailed Instructions to help you better understand the task.

If you are unable to decide or have other issues with the sample, you can **skip it**. Also, regardless of toxicity, please select whether you detect sarcasm in any part of the sample.

Detailed Instructions

**Title:** Because only money impresses women



**Comment:** He's speaking the truth though. None of you gold digging hos would date a poor man. You're lying if you say otherwise.

**1. Is the sample, as a whole, toxic?**

It is toxic.

It is not toxic.

Skip

**2. On a scale of 0-5, how confident are you about your answer?**

0 ————— 5

**3. Do you detect any sarcasm in the sample?**

Yes

No

Submit

Please watch this video for detailed navigation of the annotation platform.

(b)

Screenshot of the annotation platform, for the context-rich tasks.

### Understanding Social Media Communication

The goal of this experiment is to understand the levels of toxicity in online communication. The platform is best viewed in Desktop browsers (mobile browsers might not load correctly).  
**Content Disclaimer:** Some of the content is toxic and may cause discomfort. You can stop the annotation at any point.

**Instructions**

**Make sure you have filled the two surveys to be eligible for payment!! During this task, you need to participate only once per survey.**

- Demographics Survey
- Feedback Survey

Please read carefully all the information you have available about a sample and try to determine whether it is toxic.


The samples you will see include the title of a post, an image embodied in the post, and a comment that is related to this post. Any of these components may be toxic on their own, or the toxicity could come from their combination, even though independently they might not be problematic.

**We define toxic content as:** Any form of content that is hurtful, derogatory or obscene, such that it can hurt a person either physically or emotionally, e.g. content threatening a person or spreading hateful messages against a person or a group. Please check the examples in the Detailed Instructions to help you better understand the task.

If you are unable to decide or have other issues with the sample, you can **skip it**. Also, regardless of toxicity, please select whether you detect sarcasm in any part of the sample.

Detailed Instructions

**Title:** Because only money impresses women



**Comment:** He's speaking the truth though. None of you gold digging hos would date a poor man. You're lying if you say otherwise.

**1. Is the sample, as a whole, toxic?**

It is toxic.

It is not toxic.

Skip

**2. On a scale of 0-5, how confident are you about your answer?**

0 ————— 5

**3. Do you detect any sarcasm in the sample?**

Yes

No

**4. Is the comment toxic?**

Yes

No

**a. If you believe the comment is toxic, please highlight the part of the text that leads you to that opinion.**

You have selected:

None of you gold digging hos would date a poor man

I agree with this selection, if correctly represents the toxic content.

**5. Is any part of the context toxic?**  
(i.e. the title or the image)

Yes

No

**a. If you believe the title is toxic, please highlight the part of the text that leads you to that opinion.**

You have selected:

I agree with this selection, if correctly represents the toxic content.

**b. If you believe the image is toxic, please draw a square around the part of it that leads you to that opinion.**

Reset

Submit

Screenshot of the annotation platform, for the context-rich tasks, after the annotation is completed.

Figure 4. Our annotation platform for experiments with comments and with the additional conversational intent.

**Table 1.** The examples used to qualify workers. All the examples have been collected during the Black Lives Matter movement during the summer months of 2020

Example	Label
Get this fucking bullshit out of here you fucking Russian skills	Toxic
The guy is black. Anything over 10 cents is unreasonable. Jokes reddit	Toxic
Fuck you. America is amazing and has generated more prosperity for the world than anything. RIP Mr Floyd but fuck dumb shit-brains like you who will stand on his grave to insult America. I hope someone knocks you the fuck out	Toxic
Not saying i agree with the cops, but to be fair, protesters in Minneapolis were videotaped damaging and breaking property including cop cars. That is a significant difference in protests that would warrant a different response from police. Just wanted to say that since i haven't seen it mentioned	Not Toxic
Opinions don't matter here, you're making unsubstantiated claims about the majority. Either provide evidence or admit you could be wrong and stop giving your opinion out like its fact	Not Toxic
Not to be white. But whites are a disease on humanity. Mainly white males	Toxic
Stop referring to yourself as "black". We are fucking Brown. Stop calling us "niggas" stop calling us "black"	Not Toxic
Why the fuck would any man let a woman manage their finances?	Toxic

and the image, which led them to that opinion. Examples annotated as toxic can be seen in Figs. 3(b) and 4b.

We also note that the supplementary questions regarding the annotator rationales are only visible if the sample is considered toxic. This conditional nature of the annotation scheme was deliberate to avoid confirmation biases. In our pilot experiments, we observed that the workers might be prone to exaggerate labelling content as toxic if they see all the questions and requests a priori; thus, a set of questions focusing on the analysis of toxic parts could lead them to annotate the sample as toxic.

After the experiment, we ask annotators to optionally provide feedback. We remark that the feedback presented by all the workers in all our experiments was that they had a positive experience. Workers have praised the clarity of the instructions and how the platform facilitates the task; they judged the task as straightforward.

### 3.5 Qualification of annotators

To participate in our annotation experiments, workers first had to qualify, answering several questions that are similar to the samples in the experiments. More specifically, the initial requirement is to ensure all workers understand the adult nature of the tasks. The participants are required to be over 18 years old and have to consent to work on assignments that may include adult content.

We also wanted to establish a baseline of common understanding of the task and prime the annotators with regard to specific definition of toxicity (Section 3.1). Therefore, as quality control, we required all workers to answer whether a set of eight examples are toxic or not. These examples are provided in Table 1. Qualifications required answering at least seven out of the eight examples correctly.

We also collect demographic information for every annotator. We request all participants to complete a demographics survey and provide their socio-economic information to understand their background. Specifically, our survey consists of questions on the annotator's gender, age group, annual income range, education level, nationality, ethnicity, country of residence and political orientation. We restrict ourselves to self-reported political orientation on a simple scale

**Table 2.** Synopsis of annotations and annotation tasks. There are 3 annotations per sample, for every 500 samples (hence the sum of total annotations). All features used as controls introduce a new set of 500 samples

	Gender		Political orientation			Age		Total
	Female	Male	Left	Centre	Right	< 30yo	≥ 30yo	
Comments only	3	3	3	3	3	3	3	10,500 (21×500)
Context+Comments	3	3	3	3	3	3	3	10,500 (21×500)
Total	6000 (12×500)		9000 (18×500)			6000 (12×500)		21,000

from left to right. Most political charts, such as Nolan’s chart (Eysenck 1968) or Mitchell’s chart (Mitchell 2007) are focused on either American or European political contexts. Our goal was to use broader scales to accommodate other parts of the world. Our decision was inspired by experimental research in political philosophy by Kroh (2007) and Zuell and Scholz (2016), which has been shown to capture general trends using a 10- or 11-point scale. We ask the annotators to rate how they self-identify themselves politically, from 0 to 10 (0 being left and 10 being right). We note that the annotators cannot pass the qualification round if they have not filled the demographics survey. We use some of this demographics information to employ the controls we describe in Section 5.

#### 4. Dataset overview

In this section, we give a general overview of our dataset. Overall, our dataset consists of distinct 1,500 annotated samples collated in three batches of 500 samples. Each of the three batches is associated with a control experiment over three demographic variables – gender, age and political orientation. Each sample is annotated by three different annotators. Additionally, we have a general experiment, where for every demographic control group we examine the effect of context. Note that for each sample we have two tasks a) comment only (C) and comment with the associated context (CTM). The annotation of each of these is done by two different annotators. We emphasise that this is a between-subjects (or between groups) study design, where different annotators test different conditions, hence each person is only exposed to a single user interface. This was especially done to minimise learning and transfer effects across conditions. However, while we have taken all necessary precautions to minimise the noise, we remark that the design of the experiment (between-subject design) may indicate a potentially emphasised effects of random noise in our annotations compared to within-subject design. We also note that the visual information can either be a photo or an image that contains textual information such as memes or screenshots of conversations as shown in Fig. 2.

The experimental setup is consistent across all tasks concerning the required annotations, the rewards, the annotation limit (number of samples) on the workers and the qualifications (see Section 3.5 for more details). As shown in Table 2, each sample under each control has three annotations from three different annotators. We remark here that while three may not seem a large number, given the combination of different controls, the total number of annotations per sample (regardless of the controls) ranges between 12 to 18. Further, on average comments have 12 words (with a maximum of 379 words) while titles have 9 words (with a maximum of 127 words). Among the images we use as context, 77% contain textual information, while the rest are purely visual.

The reward for the annotations varies depending on the project type, that is the amount of information and the complexity of questions participants have to answer. We consulted past research (Founta *et al.* 2018) in order to estimate the amount for the reward and also considered

the duration of the experiments. We also iterated by taking the feedback from crowd-workers during the pilots. The final rewards across experiments are set to \$0.03 per assignment, for the simple comment-only experiments, and \$0.06 per assignment for the context-enriched ones. Finally, to ensure diversity of workers and minimise fatigue caused by long work, we put a limit of 50 samples per worker.

#### **4.1 Vote aggregation strategies**

We consider several aggregation methods. Starting with a simple majority vote (MV), that is the most commonly observed value among annotations, which is later enriched with ‘confidence in decision’ information to produce high-confidence majority votes (HCMV). These correspond to the most frequently agreed votes, taking only annotations where the annotators declared that they are very confident about the annotation (confidence score of 4 or 5, on a scale of 0–5). It is noted that, for the rest of the paper, when referring to MV and HCMV both values show the percentage of annotations that indicate toxicity. We also wanted to get an understanding of the annotators who are most confident about their annotations, therefore we calculate the rate of fully confident annotations, regardless of toxicity.

As toxicity is a very subjective concept, we also considered investigating how many of the non-toxic samples, as decided by the majority, have at least one person disagreeing. Therefore, we also calculate at least one toxic vote (ALOTV) score.

To look at sarcasm, we follow the same procedure as presented above where we calculate the majority sarcasm vote (SMV) between annotations of a sample, as well as the number of samples with at least one sarcasm vote.

Finally, we also study the inter-annotator agreement. We note that we do not expect high agreement, this is due to the prevalence of subjectivity in the task, the inherent biases we observe in our controlled experiments, and the potential random noise effects due to the between-subject experimental design. We measure agreement using  $P(A)$ , the percentage of cases where annotators agree (100% refers to full agreement) and Fleiss Kappa (Fleiss 1971), a common inter-annotator agreement metric which also takes into account the probability of agreement by chance.

#### **4.2 Summary of results**

Collectively, we have a total of 21,000 annotations for the 1500 total samples; however, it is noted that there are a varying number of annotations per sample, from 12 to 18, depending on the batch it belongs to. Samples annotated while controlling for political orientation, for example, have been examined by three different groups of workers, twice each (once without and another including the context), hence a total of 18 times.

According to the majority vote, around 30% of the samples are annotated as toxic; however, we note that nearly all samples (94%) have at least one worker suggesting they are toxic (ALOTV). Furthermore, 80% of all majority votes (and 33.5% of toxic votes in specific) are attributed with high confidence, as their annotations show a median confidence score higher than 4 out of 5. We also observe that a small percentage of samples are annotated to be sarcastic (SMV), even though the total rate of sarcasm-positive answers is more than 30%. Finally, only half of the total annotations are marked as fully confident (confidence score = 5), showing the difficulty of individuals to decide upon such a subjective matter as toxicity.

## **5. Experiments**

To better understand the various confounders influencing the perception of toxicity, we study the impact of three confounding variables. We are mainly interested in understanding the effects of

**Table 3.** Inter-annotator agreement metrics: P(A) and Fleiss' Kappa ( $\kappa$ ) for all three control experiments. Due to the different number of annotations per batch, it is not possible to calculate an overall value for the entire dataset

	Total		Non-contextual (C)		Contextual (CTM)	
	P(A)	$\kappa$	P(A)	$\kappa$	P(A)	$\kappa$
Gender	0.59	0.23	0.65	0.36	0.57	0.20
Pol. orientation	0.57	0.09	0.60	0.11	0.55	0.09
Age	0.54	0.13	0.59	0.22	0.51	0.10

demographic factors – age and gender, and political views towards the perception of toxicity. As our data is collated from Reddit forums between 2019 and 2020, it contains a sufficient number of samples that are politically motivated. This is because (a) there is an over-representation of traffic from the United States of America; and (b) a major political event in the USA – the election. We note, however, there we made no deliberate effort in filtering data that are politically relevant or express a particular view. The data collected simply reflects the popularity of topics in this period (we elucidate this in Section 3.3). We randomly sampled data among the 1500 total samples for each of our experiments below. In what follows, we describe the three main demographic control experiments and also present the results and analyses. We summarise the inter-annotator agreement for the control experiments in Table 3.

### 5.1 Controlling for gender

Recent research and surveys (Ipsos 2016; Capezza *et al.* 2017) have highlighted a consistent pattern where the perception of toxic content varies between genders. Further, practitioners who use machine learning and natural language processing for the detection of toxic online content have considered using gender as an additional feature and it has been seen to improve predictive performance in such models (Waseem and Hovy 2016). With these studies as background, we conduct a controlled experiment in order to understand the degree to which gender can potentially act as a confounding factor in perceiving toxicity.

During the preliminary analysis of data, we observed that the self-reported gender identities of our workforce were predominantly represented by males and females, all other genders were considerably under-represented on AMT. We therefore restrict our experiments to the dominant representation and use the self-reported gender to separate males from females. Even though we allow other individuals who self-report as any other gender to participate in both tasks, our final dataset consists of only the two most representative genders in these tasks, therefore we characterise the control groups hence referred to as *males* and *females*. We will release the full dataset with the all annotations on a public platform for full analysis.

Overall, the total toxicity rate of all annotations controlled for gender is 51.45%. There are 246 toxic samples, and approximately 50% of the batch is annotated as toxic according to MV. There are 12 annotations for every sample, therefore we consider the final decision of the toxicity to be fairly robust. When we consider only high-confidence annotations, most of the votes over samples remain the same and only a few change (approximately 90% remain intact). Additionally, nearly 50% of the toxic votes are attributed with high confidence (according to HCMV). We note that more than half of the annotators exhibit full confidence (5 out of 5) in their annotations.

We further note that most samples contain at least one annotation as toxic (96% of the samples with ALOTV equal to toxic). We also note that samples are not frequently annotated as sarcastic: only 15% of the samples appear to have sarcasm as the majority vote (SMV), even though the total



**Table 4.** Annotation summary of toxicity for gender-controlled experiments. The third column exhibits the rate of fully confident annotations, regardless of toxicity. The abbreviations F and M stand for Females and Males, accordingly

Task	MV (%)	HCMV (%)	Full confidence (%)	ALOTV (%)
F (comments only)	50.60	47.80	65.07	73.00
M (comments only)	47.40	43.20	50.27	71.40
F (comments+context)	55.71	46.89	49.43	82.36
M (comments+context)	51.40	43.60	56.53	81.20

**Table 5.** Annotation summary of sarcasm, for gender-controlled experiments. The third column indicates the proportion of the sarcastic samples that are also toxic (according to MV). The last column indicates the non-aggregated (for all annotations) rate of sarcasm

Task	Sarcasm MV (%)	Toxic rate (%)	Total sarcasm (%)
Females (comments only)	22.00	55.45	29.13
Males (comments only)	23.60	54.24	29.40
Females (comments+context)	37.40	68.98	40.88
Males (comments+context)	26.00	56.92	32.67

sarcasm rate of all 6000 annotations is more than 30%. We observe that half of these sarcastic samples are also toxic.

*Contextual & Control Group Findings.* Table 4 presents an overview of the aggregations related to toxicity. For the case of the simple majority vote (second column), we observe that the proportions of annotations annotated as toxic increases with the provision of context. However, when we consider annotations that are highly confident on the other hand, we do not observe any changes. We also observe that context plays an important role for *females*. Aggregated annotations by the *male* group, on the other hand, do not seem to be affected by the presence of context.

Samples that have at least one toxic annotation also consistently increase considerably in the presence of context, from around 70% to around 80%. We observe that the introduction of context also introduces subjectivity making individuals disagree more when compared to the annotations for which comments only are provided, where the tendency is to agree on no-toxicity. We also perform a Wilcoxon signed-rank test (Wilcoxon 1992) in order to assess the differences between male and female annotations by focusing on MV and annotator confidence. We consider the confidence as a continuous variable between 0–5. We observed that the annotations are significantly different among comments with context setting (with  $p < 0.05$ ) between the two groups.

Considering the summary of sarcasm annotations (presented in Table 5), we again observe that there is a marked difference between the C and CTM for *female* group for annotations of sarcasm. This seems to be reflected in all of the settings. It also seems that sarcasm is correlated with toxicity for the *female* group. None of these effects are noticed as strongly on the male group.

*Annotator Agreement.* We also compute the per cent agreement and Fleiss Kappa to obtain the inter-annotator agreement. We observe that for the whole experiment, the per cent agreement –

**Table 6.** Annotation summary of sarcasm, for political orientation-controlled experiments. The second column indicates how many of the sarcastic samples are also toxic. The last column indicates the non-aggregated (for all annotations) rate of sarcasm

Task	Sarcasm MV (%)	Toxic rate (%)	Total sarcasm (%)
Left (comments only)	23.20	40.52	27.60
Centre (comments only)	15.60	37.18	21.67
Right (comments only)	14.60	31.51	24.73
Left (comments+context)	31.00	40.65	36.27
Centre (comments+context)	17.60	47.73	26.53
Right (comments+context)	22.00	10.00	28.87

$P(A)$  – is 0.59 and within each annotation task this number rises to up to 0.65. With the introduction of context, however, we see a drop in agreement rates. This drop is even more apparent when looking at Fleiss Kappa scores –  $\kappa$ . The agreement among annotators when they only see comments is 0.36, but it decreases to 0.2 with context. Further investigating into each group, the trend is the same for all cases, both for  $P(A)$  and for Fleiss Kappa.

## 5.2 Controlling for political orientation

Previous work in the area of political science and the area of political psychology has studied the question of the influence of political orientation on the perception of both hateful language and free speech (Davis and Silver 2004; Lindner and Nosek 2009). Rossini (2019) studies the conversational dynamics when it comes to political discourse on online platforms. In this work, we are interested in understanding the differences in perception of similar types of content due to political orientation.

We recognise that it is very difficult to cover the broad spectrum of political beliefs and distinguish political orientation, especially considering that AMT workforce is spread worldwide. To avoid any definitions, we let annotators decide how they identify politically, on a scale of 0–10. Zero, in our case, indicates left-leaning political beliefs and 10 is right-leaning. We make the scale as such to provide a solid middle ground, an option for those who do not identify politically with either side and consider these annotators as a separate group. Therefore, we consider three control groups for this experiment – left-oriented, right-oriented and politically neutral annotators.

We randomly select a set of 500 samples from the 1500 samples main set, without replacement. Note that none of the samples that appear in Section 5.1 are considered for this experiment. It is interesting to observe that the rate of samples with at least one toxic annotation is at the same level as in Section 5.1 (ALOTV 94%). High-confidence toxic aggregations decreased (HCMV 18%), while fully confident annotations also decreased by a narrow margin (42.8%), and we tend to observe that annotators tend to give annotate with lower confidence. Lastly, this pattern is also consistent in answers for sarcasm (in Table 6), where sarcasm rate is fairly high (a little less than 30%) and nearly all samples have at least one individual saying they are sarcastic, yet the total set of sarcastic samples is very small (32 units or 6.5%).

*Contextual & Control Group Findings.* We present our results in Table 7, we observe that right-leaning individuals tend to overlook toxicity compared to the other groups. The tendency of left and centre is to annotate around a quarter of the samples as toxic, a rate that further increases with the introduction of context. The numbers we see in the MV column for these groups are almost double the rates of the aggregated scores discussed earlier. This shows major differences

**Table 7.** Annotation summary of toxicity, for experiments controlled about political orientation. The third column exhibits the rate of fully confident annotations, regardless of toxicity. The abbreviations L, C and R stand for Left, Centre and Right political orientations, accordingly

Task	MV (%)	HCMV (%)	Full confidence (%)	ALOTV (%)
L (comments only)	27.40	25.20	50.07	55.00
C (comments only)	23.40	14.20	35.13	56.20
R (comments only)	10.20	9.20	35.27	46.40
L (comments+context)	34.80	30.00	46.40	69.00
C (comments+context)	31.00	24.60	56.53	69.60
R (comments+context)	4.80	8.40	33.53	33.20

between the perceptions of the groups (predominantly the right-leaning v/s others). We also see that the right-leaning sub-group tends to disagree with the other groups. We also observe similar trend for the right-leaning sub-group, where we see that for the comment-only tasks we see 10.2% toxic samples are annotated as toxic. However, with conversational context (or the intent), this percentage drops to a staggering 4.8%.

In addition, for ALOTV results we find that while the total rates relatively drop compared to Section 5.1. We observe that there is a high degree of consensus between the right-leaning annotators. We also notice that the right-leaning annotators consider more than 50% of samples to be totally non-toxic which increases to approximately 70% with the added context.

Lastly, it is interesting to note the divergence in fully confident annotations. With the comment-only annotation tasks, left-leaning annotators are half the number of times fully confident about their answers, whereas the other two show much less confidence. Once the context is introduced, however, the centre-leaning group shows a sudden increase to 56.5% fully confident annotations, although the other two rates drop slightly.

With the Wilcoxon signed-rank Test on MVs, we observe significant differences in annotations across all the groups (with  $p < 0.01$ ). This is true for both the comments and comments with context settings.

Overall, these findings strongly indicate that there are extreme differences in the perception of online content between people with different political orientations, especially considering that our data originates from a platform widely used for political expression. We believe this is an important finding and expect further research in this area from the community. We also hope this finding will influence practitioners in computational social sciences to re-examine some of the experimental designs, especially in the context of automated detection of toxic online content.

*Annotator Agreement.* Regarding the inter-annotator agreement scores of the Political Orientation experiment, we see similar behaviour as Section 5.1. Agreement rates are lower than the ones from the previous experiment, but the tendency remains the same. Additionally, looking at each individual group, all scores follow the same pattern, even though there are slight differences in similarities within groups.

### 5.3 Controlling for age

Finally, with the last batch of 500 samples, we perform an experiment where we control for age. Previous research has examined the influence of age in terms of the perception of online privacy, with reports of marked difference among the age groups (Zukowski and Brown 2007).

However, there is much less research in the context of toxicity judgements regarding the influence of age. This has also been discussed in the context of recent US House Judiciary Subcommittee on Antitrust law, where senators of different age groups had a markedly different understanding of online harms.<sup>h</sup>

In this experiment, we want to investigate the influence of age towards perceiving toxicity. Our hypothesis is that younger annotators will have a different perception of toxicity. We believe there is a strong effect of the Internet culture on younger individuals, as they grew up with the Internet and social platforms incorporated in their everyday lives in an increasingly ubiquitous manner. The Internet had a principal role in the development of their belief system; hence, we expect this to have an impact on how they perceive toxicity.

In our annotation experiments for the pilot study, the study in Sections 5.1 and 5.2, we observed that a large proportion of approved annotators fell in the range of 20–30 years old (about 40%) and the range of 30–40 years old (about 42%), and smaller proportions of people below the age of 20 or above the age of 40. This observation led us to considering a split of the age group in the range below 30 and the range above 30 for our experiments to investigate the effect of age in the perception of toxic online content.

Overall, we observe a total toxicity rate of more than 30%, which means that a third of all annotations for this batch were annotated as toxic. Each sample was annotated collectively by 12 annotators, and a quarter of the batch (123 samples) were judged as toxic (MV). Once again, we notice that most of the samples have at least one annotation stating they are toxic (ALOTV 92.6%). Additionally, we also notice that half of the annotations are fully confident (49.85%). Finally, as an aggregate, there are only a few sarcastic samples (47 out of the 500, or 9.4%), although total sarcasm rate is once again relatively on the higher side (33.2%).

Our experiment suggests that the differences between the two age groups, with regard to the perception of toxicity, are seemingly weak. To further understand the dynamics, we now look at each group results separately and with reference to the effect of context.

*Contextual & Control Group Findings.* Our results are presented in Table 8. We observe that annotators with age above 30 years tend to annotate more samples as toxic, compared to their younger counterparts. We also note that they predominantly annotate with higher confidence (column 4). We also observe that the presence of context in both groups leads annotators to perceive more samples as toxic. This effect is seemingly stronger on younger participants, for whom we notice an increase in HCMV and ALOTV – from approximately 30 and 50% in the case of non-contextual tasks (C) going up to 43.6 and 72.4% when associating context (CTM), accordingly. In the case of highly and fully confident annotations, there is a large contrast among groups, regardless of the presence of context. Annotators under 30 years old tend to display low confidence. While those aged 30 or above showing the opposite – the highest confidence. The conversational context causes a slight drop in both.

Finally, as seen in Table 9, the annotation of sarcasm rises considerably with context, more so for the group of 30 and over. It is likely that this particular group has difficulties understanding sarcasm by seeing only the comments, which probably explains why there are almost double the number of samples annotated as sarcastic in the presence of the context. They also seem to associate sarcasm with toxicity more than their younger counterparts.

Based again on the Wilcoxon signed-rank test on the MVs annotations, we observe that annotators under 30 and over 30 do not show significantly different annotations (with  $p > 0.05$ ). This is true for the both the comments and comments with context settings. These observations indicate that despite the marked differences in understanding toxicity and sarcasm among the two groups,

<sup>h</sup><https://www.c-span.org/video/?474236-1/heads-facebook-amazon-apple-google-testify-antitrust-law>

**Table 8.** Annotation summary of toxicity, for age-controlled experiments. Similarly with the two previous cases, the third column exhibits the rate of fully confident annotations, regardless of toxicity. As described before, the splitting point for the groups is the age of 30; therefore, annotators of the first group are younger than 30 years old and of the second are thirty years old or over

Task	MV (%)	HCMV (%)	Full conf. (%)	ALOTV (%)
Age < 30yo (comments only)	22.20	31.20	39.73	49.40
Age > 30yo (comments only)	29.60	26.80	63.20	55.00
Age < 30yo (comments+context)	31.00	43.60	35.87	72.40
Age > 30yo (comments+context)	34.20	27.20	60.60	67.20

**Table 9.** Annotation summary of sarcasm, for age-controlled experiments. The third column indicates how many of the sarcastic samples are also toxic. The last column indicates the non-aggregated (for all annotations) rate of sarcasm

Task	Sarcasm MV (%)	Toxic rate (%)	Total sarcasm (%)
Age < 30yo (comments only)	24.20	25.62	32.27
Age > 30yo (comments only)	15.40	42.86	23.67
Age < 30yo (comments+context)	36.00	33.89	40.53
Age > 30yo (comments+context)	31.60	43.04	36.27

there are stronger differences in degrees of belief, as over 30 annotators show stronger confidence and higher cohesion.

*Annotator Agreement.* Finally, we compute the inter-annotator agreement across the Age experiment. Table 3 suggests that context results in reduced agreement among the annotators and leads to a bigger variability in terms of the annotations.

## 6. On the importance of context

In all of our control experiments in the previous section, we observe a consistent theme where we notice that the context can change the outcome of the annotation. In this section, we delve deeper into the influence of context.

We begin by combining the entire dataset over two groups – one containing all 1500 samples with only comments (C) and the other with context and comments (CTM). We compare the analysis with our previous set of observations from Section 5. We note that in this case we are merging annotations, we will have 6–9 annotations per sample.

As in the Section 5, we notice that there is a marked effect due to the presence of context on toxicity and sarcasm rates. In most of the control groups, the addition of context leads to an increase of toxic annotations, except the right-leaning political group where there is a decrease in the annotation of toxic online content. We also observe that the comment-only experiments show a total toxicity rate of 33.5% and contextual experiments show a rate of 37.9%, while the total sarcasm rate varies from 26.8% to 34.4%. We also performed Wilcoxon signed-rank test on MV's annotations, and we observed that annotations varied significantly with and without access to context (with  $p > 0.05$ ).

This pattern persists in the aggregated set of experiments, which are presented in Table 10. The increased number of annotations per sample has the effect of smoothing the noise and the

**Table 10.** Aggregated summary of annotations, for contextual versus non-contextual experiments. The first two columns show the percentage of toxic votes, as agreed by the majority and as agreed by highly confident annotations only. The same percentage is also shown for sarcasm (Sarcasm MV), along with the rate at which these sarcastic samples are also toxic. ALOTV is the percentage of samples with at least one toxic annotation. All percentages are calculated over the total number of samples

	Toxic MV (%)	HCMV (%)	ALOTV (%)	Sarcasm MV (%)	Sarcastic toxic (%)
C	28.13	31.40	76.67	10.73	34.16
CTM	31.13	36.73	88.53	16.67	41.60

marked differences between the two experiments are now much less pronounced (a result akin to the Simpson's paradox). However, in all four methods presented in the Table (columns 1–4), there is still a steep rise in the annotations of toxicity and sarcasm for the experiments with contextual information. There is also a stronger association between sarcasm and toxicity (column 5).

We also note that the difficulty to decide whether the sample is toxic in the absence of context has its frequent feedback from the annotators. This was true across workers in all of our experiments. Here is a sample feedback that we received: *'Sometimes it is difficult to know if a comment is toxic or if it was said sarcastically with just one sentence. It would be necessary to see other phrases of the conversation to know exactly in what context it was said'*.

Finally, it is not clear whether the presence of context is the main factor influencing annotator's confidence, as it shows significant fluctuations due to the characteristics of the annotators. For example, we noted in Section 5.3 that groups controlled for age show consistently strong contrasts in confidence, regardless of context. Younger participants in our experiments tend not to be very confident about their answers, whereas their counterparts show the polar opposite behaviour. We also notice that context tends to greatly affect *female* sub-group (Section 5.1) where we observe that the confidence tends to go down. We notice that the politically left-oriented and centre-oriented annotators (Section 5.2) tend to show improved confidence in the presence of context. While these changes drive our hypothesis that context can influence the confidence of annotators, however, we recognise that the pattern is fairly inconsistent. We believe that this needs further investigation.

### 6.1 Contextual toxicity

We now present our analyses on the annotations of the contextual components. As we described in Section 3.1, when annotators are provided with the context, we also request that the annotators highlight specific components in the sample that makes it toxic. Over the entire dataset, 81.2% of the toxic samples have their comment annotated as toxic and only 39.8% have the context annotated. Additionally, in more than half of the toxic annotations, it is only the comment that is found toxic (57.2% to be exact). Only 24% of cases have both comment and context toxic, and for 15.7% it is exclusively the context that is toxic. Despite the latter percentage being much lower than others, we note that it still is a substantial set of annotations which would have been considered neutral if the additional information was absent. Also, we note that titles and images played no role in the collection of the data, which was based only on comments. We expect these numbers to be reflective of real-world scenarios. Indeed, most of the state of the art in automated detection of toxic online comments, unfortunately, seem to ignore contextual cues (Vidgen and Derczynski 2020).

Distinguishing the samples among the various controls results in comment toxicity range at a similar 75–85%, indicating that our dataset is certainly rich in toxic comments. For the case of context, there seems to be a higher amount of variance. The experiments that control for gender show a rate of 52.7%, showing that more than half of the toxic annotations display a toxic bit of context, while the same rate for female exclusive group indicates it to be approximately 58.2%.



## 7. Discussion and conclusions

Conversations play an important role in our social systems and social media platforms provide a fascinating way for such interactions to happen at scale. The toxicity in such platforms is an important concern and recent work has focused on addressing and mitigating toxicity. However, a majority of previous work has studied the content without the context. Our paper studies the importance of context on the perception of toxicity. We present a novel dataset that is annotated for toxic online content, where we have focused on studying and understanding annotator perceptions of toxicity. Towards this end, we study the confounding factors along three dimensions – demographic factors such as gender and age and political orientation. While these are only some of the few observable confounders that we have studied, we envision a broader study by the community where the focus is not only on observable confounders but also latent factors that modulate the perception of toxicity. Formalising from a causal lens, these can be operationalised through techniques for estimating multiple treatments in the presence of unobserved confounding factors which can then allow for accurately estimating the causal effects even when the latent factors are misspecified (Ranganath and Perotte 2018; Wang and Blei 2019).

Our analyses highlight two key factors that can help in developing a better understanding of toxic online content. First, content on the social network platforms is contextual and ignoring the context may have significant consequences in the perception of toxicity. Second, given the subjective nature of the task, annotations of toxic online content must always consider looking for potential confounders and the final dataset should probably readily include this in the experimental design.

Recent work in natural language processing has focused on studying the risk of biases, including racial and gender biases in datasets that study toxic online content (Sap *et al.* 2019a). Our study indicates that such biases potentially exist as confounding variables and emerge in annotations. Our study further indicates that we have to be careful while considering the annotated labels as 'ground truth' for training supervised machine learning models. Majority vote perhaps doesn't reflect the true label and explains away the uncertainty in human annotations. This may have unfortunate annotation artefacts in the dataset. Through our analyses in Sections 5 and 6, we urge the community to carefully consider the experimental design and take into consideration both the observable and hidden confounding factors when using such datasets for building predictive models.

More broadly, our analyses over the dataset reveal that the datasets have enormous potential to answer what makes toxic content perceived as such. We believe that the dataset would be of significant interest not only for communities of computational social sciences and machine learning practitioners who work on toxic online content and online harms but also to researchers in the field of political science and political philosophy.

**Author contributions statement.** PM: Conceptualisation, Methodology, Software, Validation, Investigation, Data curation, Original draft preparation, Supervision, Funding acquisition; AF: Data curation, Investigation, Visualisation, Original draft preparation; Software; LS: Funding acquisition.

## References

- Bamman D., O'Connor B. and Smith N.A. (2012). Censorship and content deletion in chinese social media. *First Monday* 17(3-5).
- Barker K. and Jurasz O. (2019). Online harms white paper consultation response.
- Basile V., Bosco C., Fersini E., Nozza D., Patti V., Rangel Pardo F.M., Rosso P. and Sanguinetti M. (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, Minneapolis, Minnesota, USA. Association for Computational Linguistics, pp. 54–63.
- Baumgartner J., Zannettou S., Keegan B., Squire M. and Blackburn J. (2020). *The pushshift reddit dataset*. In Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 830–839.

- Burnap P. and Williams M.L.** (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science* 5(1), 11.
- Capezza N.M., D'Intino L.A., Flynn M.A. and Arriaga X.B.** (2017). Perceptions of psychological abuse: the role of perpetrator gender, victim's response, and sexism. *Journal of Interpersonal Violence* 36(3-4), 1414–1436. doi:10.1177/0886260517741215.
- Chandrasekharan E., Samory M., Jhaver S., Charvat H., Bruckman A., Lampe C., Eisenstein J. and Gilbert E.** (2018). The internet's hidden rules: an empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW), 1–25.
- Cheng J., Bernstein M., Danescu-Niculescu-Mizil C. and Leskovec J.** (2017). *Anyone can become a troll: causes of trolling behavior in online discussions*. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1217–1230.
- Chung Y.-L., Kuzmenko E., Tekiroglu S.S. and Guerini M.** (2019). *CONAN - COUNTER NARRATIVES THROUGH NICHE SOURCING: A MULTILINGUAL DATASET OF RESPONSES TO FIGHT ONLINE HATE SPEECH*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. Association for Computational Linguistics, pp. 2819–2829.
- Danescu-Niculescu-Mizil C., West R., Jurafsky D., Leskovec J. and Potts C.** (2013). *No country for old members: user lifecycle and linguistic change in online communities*. In Proceedings of the 22nd International Conference on World Wide Web, pp. 307–318.
- Davidson T., Bhattacharya D. and Weber I.** (2019). Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv: 1905.
- Davidson T., Warmley D., Macy M. and Weber I.** (2017). Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv: 1703.04009.
- Davis D.W. and Silver B.D.** (2004). Civil liberties vs. security: public opinion in the context of the terrorist attacks on america. *American Journal of Political Science* 48(1), 28–46.
- Djuric N., Zhou J., Morris R., Grbovic M., Radosavljevic V. and Bhamidipati N.** (2015). *Hate speech detection with comment embeddings*. In Proceedings of the 24th International Conference on World Wide Web, pp. 29–30.
- Eysenck H.J.** (1968). *The Psychology of Politics*, vol. 2. Transaction Publishers.
- Faris R., Ashar A., Gasser U. and Joo D.** (2016). *Understanding Harmful Speech Online*. Berkman Klein Center Research Publication.
- Fleiss J.L.** (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5), 378–382.
- Founta A.-M., Djouvas C., Chatzakou D., Leontiadis I., Blackburn J., Stringhini G., Vakali A., Sirivianos M. and Kourtellis N.** (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. arXiv preprint arXiv:1802.00393.
- Gao L. and Huang R.** (2017). *Detecting online hate speech using context aware models*. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria. INCOMA Ltd, pp. 260–266.
- Gomez R., Gibert J., Gomez L. and Karatzas D.** (2020). *Exploring hate speech detection in multimodal publications*. In The IEEE Winter Conference on Applications of Computer Vision, pp. 1470–1478.
- Ipsos M.** (2016). Ofcom: attitudes to potentially offensive language and gestures on TV and radio.
- Jacobs A.Z., Blodgett S.L., Barocas S., Daumé H. III and Wallach H.** (2020). *The meaning and measurement of bias: lessons from natural language processing*. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 706–706.
- Jurgens D., Chandrasekharan E. and Hemphill L.** (2019). A just and comprehensive strategy for using nlp to address online abuse. arXiv preprint arXiv:1906.
- Kiela D., Firooz H., Mohan A., Goswami V., Singh A., Ringshia P. and Testuggine D.** (2020). The hateful memes challenge: detecting hate speech in multimodal memes.
- Kroh M.** (2007). Measuring left–right political orientation: the choice of response format. *Public Opinion Quarterly* 71(2), 204–220.
- Kumar R., Ojha A.K., Malmasi S. and Zampieri M.** (2018). *Benchmarking aggression identification in social media*. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 1–11.
- Kumar S., Cheng J. and Leskovec J.** (2017). *Antisocial behavior on the web: Characterization and detection*. In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 947–950.
- Lindner N.M. and Nosek B.A.** (2009). Alienable speech: ideological variations in the application of free-speech principles. *Political Psychology* 30(1), 67–92.
- Liu P., Guberman J., Hemphill L. and Culotta A.** (2018). Forecasting the presence and intensity of hostility on instagram using linguistic and social features. arXiv preprint arXiv:1804.06759.
- Ljubešić N., Fišer D. and Erjavec T.** (2019). *The frenk datasets of socially unacceptable discourse in slovene and english*. In International Conference on Text, Speech, and Dialogue. Springer, pp. 103–114.
- Madukwe K., Gao X. and Xue B.** (2020). *In data we trust: a critical analysis of hate speech detection datasets*. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 150–161.

- Malmasi S. and Zampieri M.** (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence* 30(2), 187–202.
- Mitchell B. P.** (2007). *Eight Ways to Run the Country: A New and Revealing Look at Left and Right*. Greenwood Publishing Group.
- Nobata C., Tetreault J., Thomas A., Mehdad Y. and Chang Y.** (2016). *Abusive language detection in online user content*. In Proceedings of the 25th International Conference on World Wide Web, pp. 145–153.
- Pavlopoulos J., Sorensen J., Dixon L., Thain N. and Androutsopoulos I.** (2020). *Toxicity detection: does context really matter?*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics, pp. 4296–4305.
- Poletto F., Basile V., Sanguinetti M., Bosco C. and Patti V.** (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55, 477–523.
- Qian J., Bethke A., Liu Y., Belding E. and Wang W.Y.** (2019). A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.
- Ranganath R. and Perotte A.** (2018). Multiple causal inference with latent confounding. arXiv preprint arXiv:1805.08273.
- Rossini P.** (2019). Disentangling uncivil and intolerant discourse in online political talk.
- Sap M., Card D., Gabriel S., Choi Y. and Smith N.A.** (2019a). *The risk of racial bias in hate speech detection*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. Association for Computational Linguistics, pp. 1668–1678.
- Sap M., Card D., Gabriel S., Choi Y. and Smith N.A.** (2019b). *The risk of racial bias in hate speech detection*. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1668–1678.
- Shen Q. and Rose C.** (2021). *What sounds “right” to me? experiential factors in the perception of political ideology*. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1762–1771.
- Vidgen B. and Derczynski L.** (2020). Directions in abusive language training data: garbage in, garbage out. *PLoS One* 15(12), e0243300. arXiv preprint arXiv: 2004.
- Vidgen B., Harris A., Nguyen D., Tromble R., Hale S. and Margetts H.** (2019). *Challenges and frontiers in abusive content detection*. In Proceedings of the Third Workshop on Abusive Language Online. Florence, Italy: Association for Computational Linguistics, pp. 80–93.
- Voigt R., Jurgens D., Prabhakaran V., Jurafsky D. and Tsvetkov Y.** (2018). *Rtgender: a corpus for studying differential responses to gender*. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
- Wang Y. and Blei D.M.** (2019). The blessings of multiple causes. *Journal of the American Statistical Association* 114(528), 1574–1596.
- Warner W. and Hirschberg J.** (2012). *Detecting hate speech on the world wide web*. In Proceedings of the Second Workshop on Language in Social Media, pp. 19–26.
- Waseem Z.** (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, pp. 138–142.
- Waseem Z. and Hovy D.** (2016). *Hateful symbols or hateful people? predictive features for hate speech detection on twitter*. In Proceedings of the NAACL Student Research Workshop, pp. 88–93.
- Wilcoxon F.** (1992). Individual comparisons by ranking methods. In *Breakthroughs in Statistics*. Springer, pp. 196–202.
- Wulczyn E., Thain N. and Dixon L.** (2017). *Ex machina: personal attacks seen at scale*. In Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399.
- Yin D., Xue Z., Hong L., Davison B.D., Kontostathis A. and Edwards L.** (2009). *Detection of harassment on web 2.0*. In Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7.
- Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N. and Kumar R.** (2019). *Semeval-2019 task 6: identifying and categorizing offensive language in social media (offenseval)*. In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 75–86.
- Zampieri M., Nakov P., Rosenthal S., Atanasova P., Karadzhov G., Mubarak H., Derczynski L., Pitenis Z. and Çöltekin c.** (2020). *SemEval-2020 Task 12: multilingual offensive language identification in social media (OffensEval 2020)*. In Proceedings of SemEval.
- Zuell C. and Scholz E.** (2016). 10 points versus 11 points? effects of left-right scale design in a cross-national perspective. *ASK. Research and Methods* 25(1), 3–16.
- Zukowski T. and Brown I.** (2007). *Examining the influence of demographic factors on internet users’ information privacy concerns*. In Proceedings of the 2007 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, pp. 197–204.