

## EXAMINING DISAGREEMENT IN HATE SPEECH DETECTION: CONTEXTUALISING GENDER AND SEXUALITY

Paula Reyero Lobo, PhD Research Student, KMi

### 1. CONTACT DETAILS

#### Main contact

Paula Reyero Lobo

[paula.reyero-lobo@open.ac.uk](mailto:paula.reyero-lobo@open.ac.uk)

PhD Research Student

Knowledge Media Institute, STEM

+44 (0) 7796756117

#### Alternative Contact

Miriam Fernandez

[miriam.fernandez@open.ac.uk](mailto:miriam.fernandez@open.ac.uk)

Professor of Responsible AI

Knowledge Media Institute, STEM

You are invited to take part in a research study. Before you decide whether to take part, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully.

A challenge when annotating hate speech is identifying frequently targeted communities. This is due to the use of specialised, nuanced language.

Individuals or groups from target communities are discriminated based on characteristics such as their gender, sexuality, race, religion, age, or disability.

If you choose to participate, you will be required to grant consent at the beginning of the survey. You are free to withdraw at any time during your participation and without giving a reason.

In the event you feel triggered or possibly harmed by viewing potentially toxic or hateful remarks and recalling unfavourable prior encounters with toxic online comments, you may consider seeking support from the following assistance services:

Stonewall

<https://www.stonewall.org.uk/>

The Information Service provides information and signposting support to the LGBTQ+ community and allies.

<https://www.stonewall.org.uk/contact-stonewall%E2%80%99s-information-service>

Q:alliance

<https://www.qalliance.org.uk/>

Q:alliance is a registered charity that provides support, information and representation for the LGBTQ+ people.

## 2. INVITATION PARAGRAPH

### Content warning

*Risks associated with this study include feeling triggered or possibly harmed by viewing potentially toxic or hateful remarks and recalling unfavourable prior encounters. You may consider the following support points:*

*[Stonewall](#), Q:[alliance](#).*

### General instructions

In this study, you are asked to categorise messages potentially containing hateful, toxic, aggressive, or abusive language.

First, identify language referencing gender and/or sexuality by specifying the exact words in the message; this applies whether the message is hateful or not.

Second, indicate whether the message contains hate speech, and if it is targeted at gender and/or sexuality.

For example;

- ***It's been 3 months and these words have only proven to be true. Suck it Butch who can't deadlift [1].***

*Gender: Not-referring. Sexuality: Homosexual. Words in the message: Butch.*

- ***Is this that bimbo twat with fish lips and hideous full sleeve tats? [2]***

*Gender: Woman. Sexuality: Not-referring. Words in the message: Bimbo, Fish lips.*

### Consent and compensation

On this page, you are shown the [full information sheet](#) and are asked to give consent. You will be compensated £7 per hour of your time.

### Task details

You must complete the study without additional support. You will categorise 15 short messages, take a 7-day break, then categorise another 15 messages. A completion code will be given only after completing the two phases. The estimated time for each phase is 30 minutes.

All answers provided are being recorded. We encourage to share your thoughts on a designated page, which will surely improve the quality of the study.

### Privacy statement

This survey does not gather personally-identifying information such as your name, email address, or IP address. Your data or response is saved in an encrypted, password-protected format. To help safeguard your privacy, please do not reveal any information that could be used to identify you. The findings of the study will only be used for scholarly purposes.

[1] Butch is a term that traditionally refers to a lesbian who appears "masculine" or acts in a "masculine" manner.

[2] Bimbo is a slang term for a conventionally attractive, sexualized, naïve, and unintelligent woman.

By ticking the "I agree" box, I confirm that I have read and understood the information above and agree to participate voluntarily in this study, and I am at least 18 years of age.

I Agree

I Don't Agree

\*If you want to understand more about this study, please contact Paula Reyero Lobo (Main Contact) or Miriam Fernandez (Alternative Contact) at the email or phone number listed above.

### 3. GENERAL INFORMATION ABOUT THE RESEARCH STUDY AND COLLECTED RESEARCH DATA

- The purpose of the research: This study aims to determine whether providing additional support of specialized language used by common hate targets influences how individuals annotate or identify references to these groups in hate speech messages.
- Ethics committee review and favourable opinion/agreement gained for the study (where the project has required formal review by HREC), for example, the following statement could be used: "This research project has been reviewed by, and received a favourable opinion, from The Open University Human Research Ethics Committee – HREC reference number: **HREC/4802/Reyero-Lobo.**"
- Type of research intervention: The online survey is the project's primary intervention. Participants' coding, rationale, and findings will be utilised to inform the annotation/coding of the dataset for machine learning algorithms for hate speech detection.
- How long the study will run: The study will run from 11/09/2023 - 31/12/2023

### 4. WHAT WILL I BE ASKED TO DO IF I AGREE TO TAKE PART?

In this study, we ask that you label the messages using the following codes.

First, we ask you to identify references to:

- **Gender:** *if it mentions or is about men, women, nonbinary, other genders not specified in the previous labels (other gender)\*.*
- **Sexuality:** *if it mentions or is about heterosexual, homosexual, bisexual, other sexualities not specified in the previous labels (other sexuality)\*.*

- **Unclear:** if you are uncertain about its meaning and cannot determine if it refers to gender or sexuality.
- **None of the above:** if you are certain about its meaning and can determine it is not related to neither gender nor sexuality.

\* You may select all labels for those messages that refer to these identities broadly, that is, which are not specific to a particular gender or sexuality identity.

Second, we ask you to rate the "hatefulness" of the message, and if targets gender and/or sexuality. A message is considered:

- **Hate speech:** if it contains bias-motivated, hostile, malicious language targeted at a person/group because of their actual or perceived innate characteristics.

### Content Warning

You are about to read comments from Youtube, Reddit, Gab, and Twitter from existing Hate Speech and Abusive Language datasets.

The messages may contain or not a diverse typology of harmful expressions, which may independently reflect gender and sexuality in nuanced ways.

## 5. HOW WILL THE DATA I PROVIDE BE USED?

- The data will be collected on a secured server through an internal online survey website. All data is kept in accordance with The Open University's policies.
- We will analyse participants responses as to whether a hate speech message is or is not related to gender or sexuality to determine how their familiarity with these groups may influence their coding of the message, the discrepancies, and the patterns in their decisions. We will also compute their reliability and validity coding across participants by calculating the inter-rater agreements and comparing the kappa for each group.
- No personally identifying information will be recorded, and participants will be recruited from a popular crowdsourcing platform, Prolific. Therefore, we will not have any access to any personal information except their prolific IDs.
- We will share research findings with The Open University in publications and potentially in the media. All data will be de-identified.

## 6. YOUR RIGHT TO WITHDRAW FROM THE STUDY

- We will not be able to provide you with the option to opt out of the study once you have completed the survey since we are not gathering any personal information and your entries are anonymous.

## **7. HOW DO I AGREE TO TAKE PART?**

Participants must be at least 18 years and above to participate in the study. Be a registered participant on the popular crowdsourcing platform Prolific.

## **8. THANK YOU**

Thank you for your time and your participation.

## **9. DATA PROTECTION**

The Open University is the Data Controller for the personal data that you provide.

The lawful reason for processing your data will be that conducting academic research is part of The Open University's public task. (The consent we request from you relates to ethical considerations)

You have a number of rights as a data subject:

- To request a copy of the personal data we have about you
- To rectify any personal data which is inaccurate or incomplete
- To restrict the processing of your data
- To receive a copy of your data in an easily transferable format (if relevant)
- To erase your data
- To object to us processing your data

If you are concerned about the way we have processed your personal information, you can contact the Information Commissioner's Office (ICO). Please visit the ICO's [website](#) for further details.

If you have any queries or questions about taking part, contact the lead researcher of the study here: [paula.reyero-lobo@open.ac.uk](mailto:paula.reyero-lobo@open.ac.uk) (more information in the header of the document)