ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240957

Enhancing Hate Speech Annotations with Background Semantics

Paula Reyero Lobo*, Enrico Daga, Harith Alani and Miriam Fernandez

Knowledge Media Institute, The Open University

Abstract. Most automated hate speech detection models rely on human annotations for training and evaluation. Logic and research indicate that people who belong to groups targeted by hate speech are better at identifying it, often due to their increased familiarity with the topic and associated hate speech terminology. However, most hate speech annotation practices overlook this issue, and hence the labels produced tend to have a reduced accuracy. In this paper, we describe an approach where the text to be annotated is supplemented with background semantics, to expose the meaning of hate speech terminology that is less likely to be known to general annotators. We test the impact of this approach by measuring change in inter-annotator agreement, before and after introducing semantics, between two groups of annotators; those who belong to the target group of hate speech, and those who are not. Our experiments show that infusing text with semantic background increases inter-annotator agreement by up to 11.3% on average, aligning the annotations from annotators who do not belong to the target groups with those from the target groups.

1 Introduction

Content warning. *This research aims to tackle hate speech and contains examples of triggering and harmful language.*

Hate speech is a growing problem in societies, fuelled by the increase of online communication platforms, and the decrease in moderation efforts and their effectiveness. The recent European Commission Digital Services Act[1] is one example of regulations that aim to curb the rise in online hate speech and demands online platforms to ensure the accuracy of their automated content moderation methods. However, current hate speech detection methods are prone to error and their AI models can be inherently biased and discriminatory [11].

Recent research highlighted the increased necessity of improving the core annotation process that underlies hate speech detection models [4]. Most relevant literature uses majority votes to process annotations [26, 35]. However, majority voting is problematic since it could strengthen biases and thus reduce the overall accuracy of annotations [1]. The problem is exacerbated by the common use of offensive slang terminology in online hate speech that is often unfamiliar to the majority of people. For example, a social media post with the word *"sheboons"* could be perceived as non-hateful, or the target of hate might not be easily recognised, by individuals who are unaware of the derogatory meaning of the term which targets Black Women [37]. As a result, hateful content could be mislabelled during conventional human annotation tasks, which in turn affects AI models.

People's perception and ability to identify hateful content is often influenced by their own characteristics, experiences, and topic familiarity [24]. To this end, recent research emphasised the need to involve people who are regular targets of hate speech in the data collection and annotation processes [31, 33]. To obtain reliable annotations, it is important to cater for potential ambiguity or misunderstanding experienced by annotators who are less familiar with particular hate speech [19].

In this paper, we experiment with supplementing the text to be annotated with background semantics, to expose the meaning of hate speech terminology that could be less familiar to most annotators. Semantics was used as background information in previous research to support the annotation of domain-specific content, performed by domain experts (e.g. medical experts annotating medical reports) [7]. However, in our work, we focus on providing background semantics about individuals or groups that are targets of hate, and assess the effectiveness of this approach by comparing inter-annotator agreement before and after introducing semantics between two main groups of annotators; those who are in the hate speech target group, and those who are not, such as in the example below:

Original text: ''*the kebabs are a bunch of homosexuals*"
A post with background semantics: ''*the kebabs are a bunch of homosexuals* "
kebab: ethnic slur [...]
homosexual: sexual attraction to [...]

Our hypothesis is that making the annotators more aware of the meaning of certain terms may change their perspectives of whether a post is hateful and could help them identify the targets of hate speech more accurately. To test this hypothesis, we design a study where a diverse set of people (e.g., men, women, non-binary) annotate social media posts with and without the addition of semantics, and study the change in their annotation outputs. We focus on gender and sexuality content and groups due to them being frequent targets of hate speech. Nevertheless, our methodology is not limited to these groups.

In particular, we aim to answer the following primary questions:

1. How does supplementing text with semantics affect hate speech annotation agreement?
2. How do semantics impact convergence between annotators from hate speech target and non-target groups?
3. How does the identification of hate target groups change after introducing semantics?

* Corresponding Author. Email: paula.reyero-lobo@open.ac.uk
[1] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Our evaluation results demonstrate that adding semantic background increases inter-annotator agreement by 11.3% on average. When incorporating semantics, the annotators were more able to identify hateful content and its target groups. In addition, annotations from annotators not belonging to a hate target group became more positively correlated to the ones provided by annotators belonging to the targeted group. For example, annotations from heterosexual cisgender men became more aligned to men of gender and sexual minorities (e.g., transgender, homosexual) once semantics were added.

The remainder of the paper is organised as follows. Section 2 positions our work from the literature. Section 3 presents the methodology of our proposed semantic-enhanced annotation. Our evaluation is presented in Section 4. Section 5 discusses our approach and findings, and Section 6 concludes the paper.

2 Related Work

Supervised approaches for hate speech detection rely on human annotations to learn to identify whether textual content is hateful, the severity of the hate speech, and whether specific groups are being targeted within such content based on their gender, race, religion, ethnicity, sexual orientation, or other sensitive attributes [9, 13].

There is however an intrinsic complexity in defining and identifying hate speech, as this is prone to subjective interpretations. This results in a sparsity of resources and benchmark corpora to train detection systems, each reflecting a subjective perception [26, 35].

Multiple approaches have emerged in the last few years aiming to improve the annotation processes behind hate speech detection systems, and other ML tasks, by addressing the problem of human label variation [25, 1].

Some works are moving away from the notion of majority voting and the use of one unique ground truth label or category, to embrace and understand disagreement between annotators. For example, [6] propose the use of annotator embeddings to model the diverse perspectives of annotators. To train ML models based on this data, they consider each annotation as a separate example (i.e, different labels may exist for the same textual content). [12] propose a method called disagreement deconvolution, which transforms classification metrics to reflect the underlying distribution of labels. When computing evaluation metrics such as precision, etc., instead of comparing each prediction to a single ground truth, they compare each prediction to multiple different ground truths, one for each annotator's label.

Other works are focused on increasing the heterogeneity of the annotators, particularly on the inclusion of communities and groups that are frequent targets of hate as part of the annotation process. E.g., [18] proposed an annotation study that included black men, black women, white men, and white women on the identification of hateful instances of misogynoir (a specific type of intersectional hate directed against Black Women). Factors such as being a victim of abuse, ethnicity, racial beliefs, context, gender, and lived experiences may influence people's perceptions and interpretations of hate speech and their labelling behaviour [31, 33].

Based on the observation that hate speech is contextual, some works have focused on providing annotators with additional information (e.g., the comments leading or following a social media post, or details of the author of such post) to help annotators make more informed decisions. The use of this contextual information has been shown to increase inter-annotator agreement and to improve the performance of related hate-speech detection models [21, 28, 18, 20].

Work on improving annotation guidelines has also been considered to reduce disagreements and ambiguities during hate speech annotation. Note that definitions of hate speech also vary across studies and annotation guidelines, hindering the generalisation of hate speech [38]. Enhancements in annotation guidelines include: asking annotators to reflect on the reasoning behind their annotations [23], the provision of definitions and examples [36], the support of training prior to the annotation task [15], providing clarifications towards the inclusion or exclusion of edge cases [38], or the use of preliminary annotation rounds to identify the nature of confusions and modify annotation guidelines accordingly [10].

Complementing previous works which have aimed at improving the annotation processes behind hate speech, we propose a novel approach in which background semantics (in the form of terms and definitions extracted from a knowledge graph) is provided to annotators as additional information to identify hateful content and whether such content specifically targets certain groups. While semantics has been used to enhance human comprehension and annotation tasks [7], to the best of our knowledge, this is the first time that the use of semantics is explored to facilitate the annotation of hate speech.

3 Methodology

This section describes: (i) how we selected and assembled the data used in our experiments (Data Preparation), (ii) the process of infusing background semantics into the data (Infusing Background Semantics) and, (iii) the designed annotation task (User Study).

3.1 Data Preparation

Human annotations used in hate speech detection tasks are typically of questionable quality due to issues such as subjectivity or lack of knowledge about how haters express themselves on social media. These characteristics of training data are particularly harmful when they affect the performance of automatic hate speech classifiers, as hate speech may go undetected or non-hateful content may be automatically deleted.

Here, we illustrate how we designed a dataset of hard-to-classify hate speech samples by identifying potential false negatives from well-established hate speech datasets (Measuring Hate Speech [32], HateXplain [23], Gab Hate Corpus [14], and XtremeSpeech [22]). These datasets contain posts from different social media platforms (i.e., Twitter, YouTube, Reddit, Gab, Facebook, and WhatsApp) with their corresponding annotations of hate speech and the targeted groups.

First, we selected a set of hard-to-classify posts, specifically focusing on those annotated with gender and sexuality groups. Hard-to-classify examples were selected based on the results obtained by the hate speech classifier outlined in [29]. Following their error analysis, we select the 3752 posts classified as targeting groups related to gender and sexual orientation (e.g., homosexuals, bisexuals), which were not identified as targeting those groups by human annotators. The observation following the error analysis is that these posts seemed to have been mislabeled by annotators as not targeting these groups.

Using the entity recognition algorithm described in [17], we linked terms within the previously selected posts with entities from the Gender, Sex, and Sexual Orientation Ontology [17]. This involved utilising properties such as `rdfs:label`, synonyms, and alternative names to establish connections between the ontology entities and the terms in the posts. To ensure a balanced representation of entities within the final dataset, we minimised the unique number of appearances of each entity while ensuring that each appeared at least once.

This refinement process resulted in a set N of 350 posts, which were potentially mislabelled in their original datasets as not targeting specific groups.

3.2 Infusing Background Semantics

In this section, we outline the construction of a Knowledge Graph (KG) consisting of concepts representing terminology commonly used to target individuals or groups in hate speech, and how we utilise it to infuse background semantics into the data sample.

Our starting point is the Gender, Sex, and Sexual Orientation Ontology [17]. While this ontology covers terminology related with gender and sexuality [29], it still lacks specific terminology prevalent in hate speech on social media platforms. To address this gap, we incorporate additional resources, including concepts from encyclopedic (DBpedia [3]) and linguistic resources (Wiktionary [37] and the English Cambridge Dictionary thesaurus [2]). Using these resources, we manually curated a KG comprising both, general hate speech terminology and terms specifically employed to target groups within our selected data sample.

We built the KG and linked it to the dataset N of 350 posts with a semi-automatic process of four steps:

- **Entity Recognition:** As mentioned in the data preparation step, we used an entity recognition algorithm [17] to link terms within the previously selected posts with entities from the Gender, Sex and Sexual Orientation Ontology.
- **Manual Pruning of Irrelevant Entities:** We pruned entities to ensure they are informative and relevant to the hate speech annotation tasks. We excluded entity recognition errors (e.g, *intersex* was extracted from posts containing the verb *is* due to having the acronym *IS*), as well as general entities (e.g., *control*). The original set of 796 entities and 350 posts was reduced to 204 pruned entities linked to 240 posts after this process.
- **Retrieval of Definitions:** We use the entity properties to retrieve an informative definition or semantic description. E.g., the post "bunch of homosexuals" has a linked entity *homosexuals* (`obo:GSSO_001591`[2]). The entity's property (`obo:IAO_0000115`) was used as a relevant definition (i.e. *the state of being sexually and romantically attracted primarily or exclusively to persons of a gender identity the same as one's own*).
- **Manual Curation of Definitions:** We extended the retrieved definitions using encyclopedic and linguistic resources. We added descriptions for 63 concepts. These were mainly related to target groups with less coverage in the original ontology (e.g., terms related to race and religion like *kebab* `wiki:kebab`[3]).

The result is a KG of 267 entities linked to a set of 240 posts. These posts are shown to the annotators with highlighted spans to the linked entities and the corresponding definitions (see Figure 1).

3.3 User Study

In this section, we outline the design of the user study and how semantics contribute to supporting hate speech annotation tasks. The study underwent formal review and obtained approval from the Human Research Ethics committee of the authors' institution.

[2] `obo`=http://obofoundry.org
[3] `wiki`=https://en.wiktionary.org/wiki

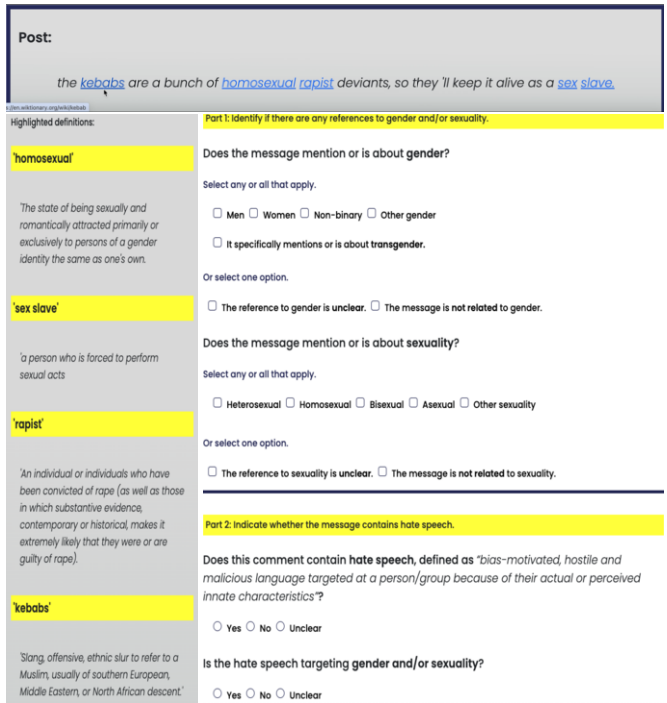

Figure 1. A semantically enriched hate speech text. Annotation involves the identification of gender and sexuality groups (Part 1) and hate speech (Part 2).

Participants 96 participants were recruited using the Prolific crowdsourcing platform [27]. Participants were compensated £7,00/hr and took an average time of 60 minutes to complete the study. Four participants were excluded and replaced for not following the instructions and failing attention checks. Participants self-identified with different genders and sexual orientations. We considered gender and sexuality minorities (i.e. non-heterosexual and non-cisgender) as the target groups, and obtained participants from those backgrounds by specifying those categories in the Prolific platform. The following distribution was obtained:

G 25% self-identified as transgender or outside a gender binary (15.6% non-binary,[4] 6.3% trans women, and 3.1% trans men),
S 25% self-identified as as being non-heterosexual (14.6% bisexual, 8.3% homosexual, 1% asexual, and 1% pansexual),
M 25% self-identified as heterosexual cisgender men,
W 25% self-identified as heterosexual cisgender women.

Sessions We designed a two-stage study to isolate the effect of background semantics following a well-established practice [39]. Posts were assigned randomly and incrementally, ensuring a balanced distribution across the G, S, M, and W groups. In Phase 1, participants were asked to annotate the posts without the support of semantics. In Phase 2, they were asked to annotate the same posts, this time with the additional aid of semantics. Sessions were conducted with one week between phases to reduce the effect of remembering decisions of the previous phase.

Both the targeted groups (G and S) and the non-targeted groups (M and W) annotated the 240 posts, with each participant annotating 15 posts. The aim was to obtain per post a minimum of three annotations from target groups, and three annotations from non-target groups (see Table 1). However, considering that the allocation of an-

[4] Participants used non-binary, queer, bigender, fluid, genderless, agender, demiman, and questioning to describe their gender.

Table 1. The user study design, in numbers.

	Participants	Posts	Annotations
Total	96	240	(96*15) 1440
(G,S)	48	15 each, 3 distinct	(3*240) 720
(M,W)	48	15 each, 3 distinct	(3*240) 720

notations was randomised, it was not possible to obtain a perfect distribution. Instead, 13% (31 posts) had either five (16), seven (14) or eight (1) annotations. In total, considering both phases of the experiment, we collect ∼3k annotations (2880).

Survey Questionnaire The questionnaire (see Figure 1) was carefully designed after consultation with two academic researchers from gender and sexuality backgrounds frequently targeted by hate speech, and after performing a pilot study involving 16 Prolific users.

Participants were asked to read each social media post and answer a set of questions. Figure 1 shows the questionnaire as it appears for Phase 2 of the study (with semantics). In Phase 1, the same layout is presented but without underlined terms in the post and with an empty column on the left. The questionnaire is divided into two main parts. In Part 1, we asked two questions: (i) if the post is about gender (About gender?), and (ii) if the post is about sexuality (About sexuality?). We consider that the post is about gender if users select one or more gender labels (men, women, non-binary, other gender, transgender). We consider the post to be about sexuality if users select one or more sexuality labels (heterosexual, homosexual, bisexual, asexual, other sexuality). In each question, users can select as an alternative that the reference to gender/sexuality is unclear, or that the post is not related to gender/sexuality. When any gender label is selected, users have to also report the exact words from the post that mention gender. The same applies to sexuality. In Part 2, we focus on hate speech. Participants are asked: (i) if the post contains hate speech (Hate speech?) and, (ii) if the hate speech is targeting gender or sexuality groups (Hate speech targeting?). For those two questions, users could answer yes, no or unclear. The "unclear" category was included in all questions to record indecisiveness. Participants received a full description of the task with examples in the Participant Information Sheet [30] and were encouraged to provide feedback via an open-ended question that they could use at any time in both phases of the study.

4 Evaluation

This section presents the analysis and results of the three questions raised in Section 1. Each subsection answers a research question. First, we measure the inter-annotator agreement before and after adding semantics to evaluate the overall impact of supplementing text with semantics in the annotation tasks (Section 4.1). Secondly, we use correlation analysis to assess the convergence of annotations from annotators not in target groups, and those in the target groups, to see this is influenced by our approach (Section 4.2). Finally, we zoom on the identification of groups targeted in given hate speech and evaluate change after introducing semantics (Section 4.3).

4.1 How supplementing text with semantics affects hate speech annotation agreement?

This section evaluates whether adding semantics as background can enhance hate speech annotations. The hypothesis is that semantics can support the annotators' understanding of the text they need to annotate, by exposing the meaning of terminology including unfamiliar slang words.

Table 2. Krippendorff's Alpha reported in phases 1 and 2 showing the difference (Δ), **boldfaced** when increased.

Gender Labels	Phase 1	Phase 2	Δ
other gender	0.242	0.087	-0.155
non-binary	0.151	0.095	-0.056
gender unclear	0.069	0.035	-0.035
transgender	0.386	0.381	-0.006
gender not-referring	0.187	0.269	**0.081**
men	0.267	0.396	**0.129**
women	0.370	0.529	**0.159**
Sexuality Labels	Phase 1	Phase 2	Δ
asexual	0.229	0.206	-0.023
sexuality unclear	0.086	0.065	-0.021
heterosexual	0.147	0.151	0.004
sexuality not-referring	0.305	0.329	**0.024**
other sexuality	0.202	0.254	**0.053**
homosexual	0.597	0.654	**0.056**
bisexual	0.213	0.295	**0.082**
General Questions	Phase 1	Phase 2	Δ
Hate speech?	0.318	0.321	0.003
Hate speech targeting?	0.255	0.260	0.005
About sexuality?	0.370	0.409	**0.039**
About gender?	0.211	0.396	**0.186**
average	0.256	0.285	0.113

We will use the inter-annotator metric Krippendorff's Alpha (α) [16], given its design and superiority in dealing with incomplete data (i.e., where not all participants annotate each post) and a variable number of raters for each post. Other metrics, such as Cohen Kappa [5] or Fleiss' Kappa [8] are less tuned to these conditions since they require a fixed number of raters. Krippendorff's $\alpha \in [-1, 1]$, where $\alpha = -1$ indicates full disagreement, $\alpha = 0$ no agreement beyond chance, and $\alpha = 1$ perfect agreement. In our analysis, we use the same levels proposed by Cohen [5], where $\alpha \leq 0$ indicates *no agreement*, and 0.01–0.20 as *none to slight* agreement, 0.21–0.40 as *fair*, 0.41– 0.60 as *moderate*, 0.61–0.80 as *substantial*, and 0.81–1.00 as *almost perfect agreement*.

Table 2 shows the α measures of inter-annotator agreement before adding semantics (Phase 1), and after (Phase 2), for the labels of *gender*, *sexuality*, and *general questions* about hate speech and its targets. The average α for Phase 1 (i.e. without adding semantics) is 0.256. This average increased by 11.3% to α = 0.285 when the semantic background is added to the annotated text. This demonstrates the overall positive impact of semantics on increasing inter-annotator agreement.

In both Phases, 5 out of 18 annotation tasks led to no agreement ($\alpha < 0.2$). When semantics were added, agreement decreased for 6 tasks out of 18. For *other gender*, agreement moved from *fair* to *none to slight agreement* in Phase 2, whereas for the other 5 tasks, the *no agreement* level was held in Phase 2, indicating that semantics did not influence these annotations. These annotation tasks are further discussed in Section 4.4.

We can also observe that in Phase 2, once semantics were added, inter-annotator agreement increased in 12 annotation tasks. When scrutinising Phase 2 results further, we observed that in 10 out of our 18 annotation tasks, agreement was at the "slight" level. Posts targeting *men, women* and *homosexual* only reached "moderate" or "substantial" agreement when semantics were introduced.

A high increase in agreement of over 10% is in *women* and *men* annotation tasks, and most tasks see a moderate ($5\% \leq \Delta < 10\%$) increase (*bisexual*, *gender not-referring*, *homosexual*, *other sexuality*). The others see a slight increase (*sexuality not referring*) or remain almost unchanged (*heterosexual*, *transgender*).

In the General Questions tasks, the top improvement in agreement

	Phase 1				Phase 2			
	M	W	S	G	M	W	S	G
other	0.195	-0.003	0.232	**0.287**	0.068	**0.495**	0.271	0.043
non-binary	0.121	0.076	**0.563**	0.366	0.085	-0.036	**0.271**	0.13
unclear	**0.091**	0.024	0.029	-0.074	0.029	0.012	**0.159**	-0.005
transgender	0.593	0.252	**0.65**	0.49	0.234	0.253	0.439	**0.565**
not-referring	0.187	**0.261**	0.122	0.209	0.216	0.253	**0.28**	0.262
men	0.268	0.35	**0.404**	0.256	0.378	0.356	0.454	**0.516**
women	0.337	**0.483**	0.296	0.393	0.439	0.515	0.587	**0.677**
asexual	0.247	-0.007	**0.563**	-0.016	0.247	-0.007	**0.429**	0.229
unclear	-0.037	0.008	**0.063**	-0.005	0.051	**0.123**	-0.001	0.044
heterosexual	0.229	-0.047	**0.359**	0.279	0.061	-0.005	**0.31**	0.185
not-referring	0.345	0.238	**0.463**	0.298	0.357	0.224	**0.485**	0.377
other	0.156	-0.018	**0.253**	0.191	0.035	0.392	**0.608**	0.461
homosexual	0.539	0.53	**0.878**	0.51	0.618	0.622	**0.748**	0.626
bisexual	0.16	0.047	**0.552**	0.049	0.247	0.351	**0.552**	0.353

	Phase 1				Phase 2			
	M	W	S	G	M	W	S	G
other	**0.46**	0.38	0.4	nan	0.0	**0.12**	0.06	nan
non-binary	0.28	0.11	**0.44**	nan	0.08	0.1	**0.3**	nan
unclear	0.26	0.07	**0.31**	nan	0.03	-0.04	**0.05**	nan
transgender	0.4	0.3	**0.63**	nan	0.5	0.39	**0.55**	nan
not-referring	0.24	**0.28**	0.27	nan	**0.35**	0.26	0.3	nan
men	0.24	**0.39**	0.3	nan	0.39	0.44	**0.56**	nan
women	0.44	**0.46**	0.41	nan	0.6	0.55	**0.62**	nan
asexual	0.31	0.1	nan	**0.68**	**0.6**	0.41	nan	0.37
unclear	**0.23**	0.15	nan	0.17	**0.14**	-0.01	nan	0.07
heterosexual	**0.24**	0.14	nan	0.23	**0.45**	0.11	nan	0.39
not-referring	0.33	**0.45**	nan	0.41	0.52	0.35	nan	**0.54**
other	**0.49**	0.19	nan	0.43	0.29	0.26	nan	**0.54**
homosexual	0.72	**0.83**	nan	0.66	0.79	**0.89**	nan	0.67
bisexual	**0.49**	0.45	nan	0.4	0.46	**0.64**	nan	0.33

Figure 2. Krippendorff's Alpha and Pearson's correlation coefficients in session without (Phase 1) and with (Phase 2) semantics; on annotations from target (G and S) and other (M and W) groups. Highest agreement and correlation with the target group, i.e., in not a number (nan) columns, are **boldfaced**.

in Phase 2 is seen in *gender*, which saw an increase of $\Delta = 19\%$, and *sexuality* with a 4% increase in agreement. In contrast, annotations for the two *hate speech* tasks remained largely unchanged.

Finding 1. *Adding background semantics increased agreement by* 11.3% *on average for gender and sexuality groups. However, the agreement went up in only 50% of the annotation tasks in Phase 2 when semantics were added, did not change in 15% of the tasks, and the inter-annotator agreement went down in 35% of the cases.*

4.2 How do semantics impact convergence between annotators from hate speech target and non-target groups?

Although the addition of semantics to the annotation process increased inter-annotator agreement overall, in some cases this resulted in reduced agreement levels. To better understand these cases, we investigate how annotations have changed in variance for such cases before and after the introduction of semantics.

Due to the lack of a gold standard in hate speech detection [4], it is difficult to evaluate "correctness" with standard analysis metrics, such as accuracy. Therefore, in this part of the evaluation, we explore the similarity of annotations from target groups (G and S) with those from the other groups (M and W) in both phases of the study.

First, we compare inter-annotator agreement across annotator groups (Section 3.3). Based on our findings, and the assumption that hate-speech target groups are closer to being experts, we compare similarity using Pearson's correlation [34]. The statistic is based on aggregated annotations for each gender and sexuality category. In each category, we measure the correlation of each group (i.e. M group) with the target (i.e., with G in *gender labels*, and vice versa).

Figure 2 shows that agreement for many categories increased with semantics in Phase 2. The G and S target groups reached the highest scores when provided with semantic background, since they seem to have agreed more in all categories except for two (i.e., *other gender* and *sexuality unclear*). In gender, inter-annotator agreement increased between the annotators that belong to the target group (G), who reached the highest agreement in Phase 2. In sexuality, semantics increased agreement in other groups, as it was already higher for the target (S). In general, positive correlations with target groups became stronger with semantics. Background semantics has especially brought closer the annotations from M (in *asexual*, *heterosexual*, sexuality *not-referring*, and all *gender labels* that increased agreement in Phase 2). In W, semantics aligned some categories (*asexual*, *bisexual*, *homosexual*, *other sexuality*, *women*, *men*), or left them unchanged. Especially in gender, semantics made target groups more correlated, increasing agreement within themselves.

Finding 2. *Our findings support the consideration of annotators from hate-speech target groups as more experts in the task. More specifically, when semantics increased inter-annotator agreement, it was because annotations from non-target groups aligned more closely with those from the target group.*

4.3 How does the identification of hate target groups change after introducing semantics?

In the previous section, we showed that all annotator groups (G, S, M, and W) became more aligned with the targets when adding semantics. Although this alignment of annotations increases inter-annotator agreements, a deeper investigation is required to better understand how the annotations are changing once semantics are introduced.

To conduct this analysis we grouped posts into ten different categories based on their annotations, and observed how these categories changed before and after infusing semantics. These categories are constructed based on two dimensions: (i) the level of agreement of the annotations in the post *(all agree, the majority agree, an opinion is formed -but without a majority agreement, no agreement)*, and (ii) if a decision is made on whether the post is targeting a group (*the post is targeting at least one group, the post is not-targeting any group, it is unclear*). For example, a post p with the set of annotation vectors before semantics $\vec{p_{bs}}$=$\{[women], [women], [unclear]\}$ can be categorised as *majority targeting*, as there is a majority voting for the category women, and it is targeting a group. We then look at the annotations of the same post after semantics $\vec{p_{as}}$=$\{[women], [women], [women]\}$. The category of this post has changed as *all targeting*, as all annotators selected the same label, and a decision is made that the post is targeting a group. Definitions of the ten categories and examples of annotations can be seen in Table 3. Figures 3 and 4 show the categories that emerge, and how they change after adding semantics.

As can be seen in Figure 3, semantics impacts *gender* and *sexuality* annotations differently. It is worth noting that in Phase 1, most posts were considered to be targeting gender (75% of the posts), whereas fewer were deemed to be targeting sexuality (25%). In gen-

Table 3. Two-step categorisation of posts, with examples where 3 and 6 annotators select men (g_1), women (g_2), non-binary (g_3), other gender (g_4), transgender (g_5), not-referring to gender (N), or unclear (U). Opinions are not applicable (NA) for less than 3 participants, as they would reach a majority vote.

Category	Rule	Examples	
Level 1: By agreement, if:		3 annotators	6 annotators
all	all select the same labels	[3: [g_1]]	[6: [N]]
majority	a majority select the same labels	[2: [N], 1: [g_3]]	[4: [g_1, g_4], 2: [g_1]]
opinions	at least two select the same labels, but not enough to reach a majority	NA	[3: [U], 3: [N]]
no-agreement	all select different labels	[1:[U],1:[N],1:[g_3]]	
Level 2: When a decision is made (i.e., except no-agreement), if the decision:			
targeting	involves at least one target group	[3: [g_1]]	[4: [g_1, g_4], 2: [g_1]]
unclear	was only unsure of being targeting		[3: [U], 3: [N]]
not-targeting	did not involve any target group	[2: [N], 1: [g_3]]	[6: [N]]

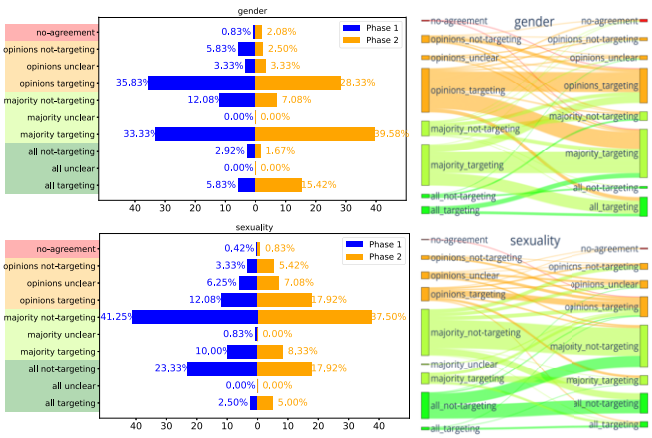

Figure 3. Frequency and category changes from phases 1 to 2. A category is assigned to each post based on its target group annotations.

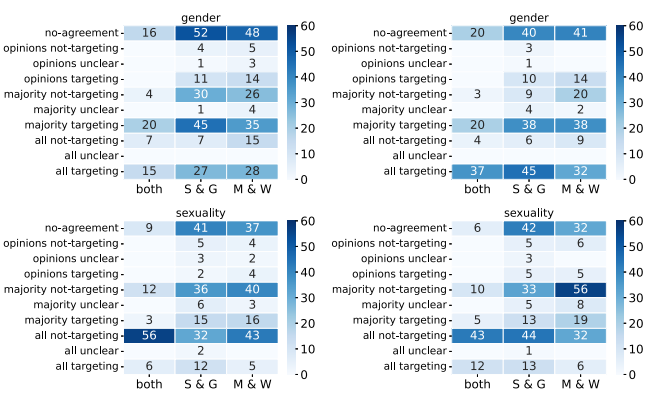

Figure 4. Hate speech categories as annotated by all (both), only target (G & S), or non-target (M & W) groups in phases 1 (left) and 2 (right).

der, more posts were identified with semantics (*all targeting* increases by 9.59%, and *majority targeting* increases by a 6.25%). Also in sexuality, more posts were identified (*all targeting* doubles, and *opinions targeting* increases by 6%). But in this case, they come from being considered as not related to sexuality. When looking at the changes between the two phases, they seem to occur in posts with low agreement (*opinions*) or only a *majority* saying is not targeting a group. In sexuality, most of the cases that become "opinionated" come from the majority labelling the posts as *not targeting* (and switched to opinions_targeting, i.e. involving targets), while the annotation cases that have already been identified in Phase 1 as having a hate-speech target, kept their label. In other words, posts labelled in Phase 1 as opinions/majority/all_targeting did not change to not-targeting in Phase 2. In *gender*, more annotators reached full or majority agreement when given semantics.

Noting the overlap between participants in hate speech categories (Figure 4), they mainly occur in cases of full agreement (i.e., clear examples). In Phase 2, overlaps mainly increased in *all targeting*, which was double as high for both gender and sexuality. A key observation is that the number of posts labelled as targeting a group was always higher for the annotators from target groups (S & G). In Phase 1, prior to adding semantics, 83 posts in *gender* and 29 posts in *sexuality* were identified by the annotators in groups S & G to have a hate speech target group, increasing to 93 and 31 respectively in Phase 2 once semantics were added (10.7% average increase). The annotators in groups M & W identified 77 and 25 such posts in Phase 1, rising to 84 and 30 in Phase 2 (11.2% average increase).

For instance, the number of posts identified by groups M & W to be targeting *gender* (all_targeting) went up from 28 in Phase 1 to 32 in Phase 2. For posts targeting *sexuality*, the number rose just slightly from 5 in Phase 1 to 6 in Phase 2. For annotators in groups G & S, all_targeting posts in *gender* went up from 27 in Phase 1 to 45 in Phase 2, and for *sexuality*, they increased from 12 to 13.

In addition to the above, we noticed that some posts were only identified as targeting a group once semantics were introduced (Table 4). In total, there were 19.6% (47 out of 240 posts) new posts identified to be targeting a gender or sexuality group in Phase 2. More precisely, 28 and 16 new posts were identified as targeting a group, by group M & W and group G & S, in Phase 2.

Addressing knowledge gaps in sexuality seemed to mostly affect non-target groups (M & W), as the number of new posts identified was visibly higher (4.58% as compared to 1.25%). The written justifications show that annotators learnt new terminology about hate speech target groups (e.g., *dike, fags* or *sheebons*); identified in some cases only by target groups (*brusband, misogynistic*, or *thot*).

Finding 3. *Adding semantics to text in annotation tasks increases inter-annotator agreement by helping to resolve hard-to-annotate text (i.e., posts with a lower agreement or initially not deemed as pointing to hate-speech target groups). In those cases, agreement increased between the annotators. A better understanding of hate-speech terminology was enabled by semantics.*

4.4 Analysing Increases in Disagreement

We examined categories where agreement decreased with semantics (see Table 2). This investigation involved analysing posts in these categories both before and after semantics were added.

Among these categories, *gender unclear* and *sexuality unclear* appeared in the highest number of posts (i.e. 115 and 140 before and after semantics). However, only 2 of the 178 unique posts containing unclear reached a majority, which indicates these categories were ambiguous and reached no agreement before or after semantics.

Table 4. Percentage, counts and terms in the posts initially considered *not targeting*, which were used to justify a target group only in Phase 2. Terms agreed by both (S & G) and (M & W) groups are **boldfaced**.

Participants	%(Counts)	Terms
		Gender
S & G	7.08%(17)	**she, her, he, his, him, bitch**, **faggot, cuck**, misogynistic, masculinity, her brusband, **gal** mama, mother
M & W	5.42%(13)	**she, her, his**, **beta cucks**, **fags**, **sheboons**, **wife beater**, **masculine**, **lady**
S, G, M, W	1.25%(3)	**her, she, cuck**
		Sexuality
S & G	1.25%(3)	foreskin, thot, married her own brother
M & W	4.58%(11)	**bi**, **dike**, **faggots**, ass, **msm (men who has sex with men)**, **her boyfriend**, **gays, lgbt**

The remaining categories appeared in 36.25% posts (i.e. 87 out of 240); i.e. 52 contain *other gender*, 52 *non-binary*, 40 *transgender*, and 20 *asexual*. In 42 posts, these categories were included in addition to the annotation agreed by most annotators. E.g., only 1 out of 6 annotators selected *other gender* in addition to women in the post "go home you dirty brown cunts". Disagreements of this type point to unresolved misunderstandings where one annotator misinterpreted the post. A key difference between categories is that the number of such misunderstandings regarding *transgender* was higher before adding semantics (i.e. 9 as compared to 3), while it remained similar in the other categories.

In the other 45 posts, more than one annotator included these categories. Posts involved multiple target groups and the less familiar ones (i.e. asexual, non-binary, transgender) were only included by a few annotators, as compared to more commonly known categories (i.e. homosexual, men, women). E.g., the post "i doubt that is even her photo or it a female" was annotated with only the category *women* by most annotators. Only three included the category *transgender* in addition to *women*. In such cases, where posts lack explicit terms indicating the additional target group, the inclusion of background semantics is not as effective, as the target remains implicit. In other posts, where background semantics could have helped to identify the additional target group, we still noticed that most annotators selected a primary target group, and disregarded the annotation of less common target groups. We also observed that adding background semantics is not sufficient in some cases, as contextual information is also needed to make a more informed decision. E.g., in the post "she is a f****** tomboy" even when providing semantic information (i.e., tomboy refers to a girl who behaves in a manner typically perceived as boyish), it is still unclear whether categories such as *non-binary* should be attached to the post without additional context.

Finding 4*. Our analysis revealed that the categories increasing disagreement with semantics were referring to groups (other gender, non-binary, transgender, asexual) that were mentioned in our social media sample only indirectly, or with less associated terminology to infuse with supporting semantic background information.*

5 Discussion

In this paper, we introduce the use of background semantics within hate speech annotation tasks by supplementing text with highlighted terms and definitions extracted from a knowledge graph. This method aims to expose the meaning of hate speech terminology that could be less familiar to most annotators, facilitating knowledge acquisition and improving text comprehension. This approach offers an innovative solution to address human label variation [25] in hate speech.

Our experiments involve groups of annotators who self-identify with various genders and sexual orientations. The results demonstrate key benefits when semantics are incorporated: (i) an increase in overall agreement among annotators, (ii) closer alignment of annotations from non-target groups with those from target groups, which tend to have greater familiarity with the subject matter and associated terminology; and (iii) higher agreement levels due to helping to resolve hard-to-annotate texts (i.e. initially not deemed as having a hate speech-target). An important step in future research is testing the impact of these improved annotations on ML model performance.

Despite the positive results obtained we also acknowledge several limitations. Firstly, the availability of background semantics covering hate speech, particularly derogatory terms towards target groups, is limited. Generating comprehensive background semantics is resource-intensive and constrained by the evolving nature of language in social media. Nevertheless, even incomplete background semantics can support the annotation of certain types of hateful content or content directed at specific groups. Besides, the inclusion of all meanings of ambiguous terms is a possible solution to scale in real-world settings.

In our social media sample, certain gender and sexuality groups were more frequently mentioned than others. A more balanced sample, including less represented target groups, could provide valuable insights into the key background semantics that could better support annotators.

In terms of participants, our recruitment conducted via the Prolific crowdsourcing platform may also introduce demographic biases. Factors, such as the cultural backgrounds in our sample, could influence annotators' decisions when labelling hate speech posts [31, 33]. Future work should explore whether our findings vary across different protected characteristics.

In terms of our designed user study, we attempted to control individual factors by assigning posts randomly and involving the same participants in both phases. However, other factors such as the order in which posts are presented to participants, and the time duration between phases, may have influenced the annotation process.

Finally, it is important to note that, our study is not aiming to resolve all disagreements in hate speech annotation tasks. Instead, background semantics aims to resolve disagreements arising due to the complexity of hate speech terminology.

Despite the above limitations, we believe our work opens a new and exciting area of research to improve the annotation process [4] underlying the generation of ML models for hate speech recognition. Improving hate speech annotations is crucial for better safeguarding individuals and communities.

6 Conclusion

This paper presented a novel hate speech annotation methodology, where the text to be annotated is enhanced with background semantics to help annotators process the hate terminology often used to target individuals or groups in social media. This approach succeeded in increasing the overall agreement and alignment of annotations with those produced by the group of annotators who are most likely to be exposed to hate speech. Our findings stress the need for reviewing existing annotated datasets for hate speech, and for adopting new annotation approaches that ensure the awareness of lay annotators to the unfamiliar terminology often used in targeted hate speech.

Supplemental Material Statement: Data and Source Code is publicly available [30].

Ethics Statement: Participants were briefed with helplines and support contact points. As defining gender/sexuality categories risks the equality of inclusion, we provide an open-ended question to record this sensitive data. This research project has been reviewed by, and received a favourable opinion, from The Open University Human Research Ethics Committee.

Acknowledgements

We thank Em Dean and Tracie Farrell for their valuable feedback on the study design, and Chris Sanders for helping to develop the online survey. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias".

References

[1] L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564.

[2] C. U. P. . Assessment. Cambridge english corpus, 2024. URL https://dictionary.cambridge.org/thesaurus/. Accessed on 03 27, 2024.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, page 722–735, 2007. ISBN 3540762973.

[4] F. Cabitza, A. Campagner, and V. Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, Jun. 2023.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

[6] N. Deng, X. Zhang, S. Liu, W. Wu, L. Wang, and R. Mihalcea. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, 2023.

[7] A. Dumitrache, L. Aroyo, C. Welty, R. Sips, and A. Levas. Dr. detective: combining gamification techniques and crowdsourcing to create a gold standard for the medical domain. In *Crowdsourcing the Semantic Web*, 2013.

[8] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[9] P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[10] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018. doi: 10.1609/icwsm.v12i1.14991.

[11] FRA. Online content moderation - current challenges in detecting hate speech. Technical report, European Union Agency for Fundamental Rights, 2023. URL https://fra.europa.eu/en/publication/2023/online-content-moderation.

[12] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[13] M. S. Jahan and M. Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, page 126232, 2023.

[14] B. Kennedy, M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, et al. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30, 2022.

[15] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[16] K. Krippendorff. Computing krippendorff's alpha-reliability. In *Citeseer*, 2011. URL https://api.semanticscholar.org/CorpusID:59901023.

[17] C. A. Kronk and J. W. Dexheimer. Development of the Gender, Sex, and Sexual Orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association*, 27(7):1110–1115, 06 2020.

[18] J. Kwarteng, G. Burel, A. Third, T. Farrell, and M. Fernandez. Understanding misogynoir: A study of annotators' perspectives. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 271–282, 2023.

[19] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, and S. Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Nov. 2021.

[20] N. Ljubešić, I. Mozetič, and P. K. Novak. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, 29(6):1481–1494, 2023.

[21] I. Markov and W. Daelemans. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying*, pages 37–42, 2022.

[22] A. Maronikolakis, A. Wisiorek, L. Nann, H. Jabbar, S. Udupa, and H. Schuetze. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, 2022.

[23] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, 2021.

[24] D. Meyer. Evaluating the severity of hate-motivated violence: Intersectional differences among lgbt hate crime victims. *Sociology*, 44(5): 980–995, 2010. doi: 10.1177/0038038510375737.

[25] B. Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731.

[26] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

[27] Prolific. Prolific: High-quality human data to deliver world-leading research and ais. URL https://www.prolific.com. Accessed on 04 04, 2024.

[28] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, and V. Cotik. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590, 2023. doi: 10.1109/ACCESS.2023.3258973.

[29] P. Reyero Lobo, E. Daga, H. Alani, and M. Fernandez. Knowledge-grounded target group language recognition in hate speech. In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, pages 1–18. IOS Press, 2023. doi: 10.3233/ssw230002.

[30] P. Reyero-Lobo, E. Daga, H. Alani, and M. Fernandez. Enhanced Hate Speech Annotations with Background Semantics - Data. 7 2024. URL https://ordo.open.ac.uk/articles/dataset/Enhanced_Hate_Speech_Annotations_with_Background_Semantics_-_Data/26212604.

[31] G. Roussos and J. F. Dovidio. Hate speech is in the eye of the beholder: The influence of racial attitudes and freedom of speech beliefs on perceptions of racially motivated threats of violence. *Social Psychological and Personality Science*, 9(2):176–185, 2018.

[32] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. von Vacano, and C. Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC*, pages 83–94, 2022.

[33] Y. Sang and J. Stanton. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer, 2022.

[34] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[35] B. Vidgen and L. Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12): e0243300, 2020.

[36] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[37] Wikimedia. Wiktionary, the free dictionary, 2007. URL https://en.wiktionary.org/wiki. Accessed on 03 05, 2024.

[38] W. Yin and A. Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

[39] L. Zhao and S. Lipsitz. Designs and analysis of two-stage studies. *Statistics in medicine*, 11(6):769–782, 1992.