Annotating Hate Speech Towards Identity Groups

Donnie Parent, Nina Georgiades, Charvi Mishra,
Khaled Mohammed, Sandra Kübler
Indiana University
{daparent,ngeorgia,chmish,mohammek,skuebler}@iu.edu

Abstract

Detecting hate speech, especially implicit hate speech, is a difficult task. We focus on annotating implicit hate targeting identity groups. We describe our dataset, which is a subset of AbuseEval (Caselli et al., 2020) and our annotation process for implicit identity hate. We annotate the type of abuse, the type of identity abuse, and the target identity group. We then discuss cases that annotators disagreed on and provide dataset statistics. Finally, we calculate our inter-annotator agreement.

1 Introduction

As online communities have become more accessible, hate speech has become more prevalent in these spaces. This surge has driven research into developing machine learning models capable of automatically detecting such harmful content (e.g., Waseem and Hovy, 2016; Zampieri et al., 2019; Warner and Hirschberg, 2012). However, existing models tend to detect hate that is explicit in nature, i.e., posts that contain explicitly hateful words, more successfully (e.g. Wiegand et al., 2019).

Implicit hate is equally widespread as explicit hate, but it often slips through the filters due to the difficulty to detect it with current systems. Implicit hate typically lacks unambiguously abusive and explicit words, has a less-predictable syntactic format, and is more affected by data bias (Lopez and Kübler, 2025; Wiegand et al., 2019).

To better detect implicit hate, Wiegand et al. (2021b) argue in favor of individual models for specific subtypes of implicit hate. They have created datasets and models for hateful comparisons (Wiegand et al., 2021a), hate towards identity groups (Wiegand et al., 2022), and euphemistic abuse (Wiegand et al., 2023). The dataset on identity hate by Wiegand et al. (2022) was created by collecting tweets where an identity group subject was followed by a negative polarity verb. Selected tweets were then annotated for hateful content. Our main goal is to determine the generalizability of this type of data, as all examples exhibit nearly identical syntactic structures. It is unclear whether such a syntactic bias has an effect on the generalizability of the data to more syntactically diverse data. For experiments to investigate this question, we need a dataset that contains implicit hate speech annotated for whether the hate targets an identity group, but without the syntactic bias. To our knowledge, such a dataset does not exist. For this reason, we have annotatad a dataset: We have extracted implicit hate speech examples from a larger dataset, AbuseEval (Caselli et al., 2020), and annotated these examples for identity hate. For a first investigation of the syntactic bias, refer to (Parent et al., 2025).

Offensive Content Warning: This report contains some examples of hateful content. This is strictly for the purposes of enabling and explaining this research. Please be aware that this content could be offensive and cause you distress.

2 Dataset

We use the AbuseEval dataset by Caselli et al. (2020), which is based on the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019). Caselli et al. (2020) retrieved only the offensive tweets from OLID and annotated them, marking each tweet containing either profanity or slurs as *explicit abuse*, any other offensive tweet as *implicit abuse*, and any non-interpretable or non-abusive tweets as *not abusive*.

Our goal is to use AbuseEval to identify hate speech against identity groups. For this reason, we selected only the tweets that Caselli et al. had annotated as *implicit abuse*, and re-annotated them. We cleaned the data, partly to increase readability for annotators, and partly to increase its usability for machine learning experiments: We replaced

Proceedings of Recent Advances in Natural Language Processing,pages 894–899
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-102

HTML tags with their respective symbols (e.g., "&" was replaced with "&"), and removed hashtags that are syntactically optional in a tweet. For syntactically relevant hashtags, we removed the "#" character and preserved any relevant information within the hashtag (such as names, acronyms, or half-words). E.g., the original tweet in example (1) was converted to the text in (2). In the example, "#ChristineBlaseyFord" is syntactically required, and we converted it to individual words. We removed "#ConfirmJudgeKavanaugh", since the sentence is well-formed without this hashtag.

(1) #ChristineBlaseyFord is your #Kavanaugh accuser... #Liberals try this EVERY time... #ConfirmJudgeKavanaugh URL

(2) Christine Blasey Ford is your Kavanaugh accuser... Liberals try this EVERY time... URL

3 Annotation

Our dataset was annotated by five annotators: four identifying as female and one as male, all with a background in linguistics. Prior to any annotation, we developed definitions of central concepts and annotation guidelines to encourage consistency between annotators (see below). Annotations were carried out in several cycles with discussions of unclear cases between cycles. Those discussions focused less on individual examples, and more on more general problems in delineating between identity abuse and other abuse, or questions concerning the use of external information. These discussions led to refinements of the annotation guidelines. We calculated inter-annotator agreement after the final round of discussions.

The annotations include three categories. We first annotate for type of abuse, with labels for identity abuse, non-identity abuse, and non-abuse. The second category, type of identity abuse, separates identity abuse into explicit or implicit references to identity group(s), (only applicable to examples of identity abuse). The third category is an open-ended category to represent the identity group(s) targeted, again only applicable for examples of identity abuse.

In the remainder of this section, we first provide the definitions that serve as the basis for the annotations, then we describe the annotation guidelines, followed by dataset statistics.

3.1 Definitions

Our annotations are based on the following definitions:

Abusive language: We follow Wiegand et al. (2019) in defining abusive language as "hurtful, derogatory or obscene utterances made by one person to another person or group of persons." Abusive language must target people to be considered abusive. Counterspeech, i.e., language that points out discrimination (such as sexism, racism, etc.) or critiques double standards regarding discrimination, is not considered abusive. Abusive language does not have to be explicit in nature. Emphatic cursing may be offensive, but is not considered inherently abusive, unless it is paired with other abusive language.

In general, definitions of abusive language tend to be too general to be easily operationalizable. Our definition of abusive language was chosen to encompass as much abusive language as possible without being overly specific. Examples of abusive language include:

(3) How is she hiding her ugly personality? She is the worst.

(4) Liberals are so pathetic!

(5) #Dems are disgusting. They only believe women who accuse conservatives of abuse crimes. When liberals are accused, they ignore the accusations.

(6) Conservatives do NOT represent all Canadians. They are Authoritarian scum.

Implicit abuse: We follow Lopez and Kübler (2025) and define implicit abuse as follows: "Instances where a comment/post alludes to stereotypes or other negative attributes and associates them with a particular group of individuals, especially as it relates to attributes related to a group or individual's identity (e.g., ethnic heritage, nationality, gender, sexual orientation, religion, disabilities, body shape)." Examples of implicit abuse include:

(7) White mask privilege at its finest. Always trying to keep a black mask down

(8) I hope d human rights rats r happy!hope Sidhu is happy! I hope d liberals r happy! Let's play cricket wth thm let's thank thm 4 wht they r doing 2our soldiers Let's celebrate our impotence funny a piddly little

nation like pak is taking us 4 a ride

(9) You know what is tiresome? Bernie supporters bashing Hillary Clinton. She is an actual Democrat who raises money for the Democratic Party and doesn't bash them. She has a positive message. So who is dividing the left?

(10) 4 out of 10 British people are basically full-on racists. 4 out of 10 voters vote for the Conservatives. Coincidence!???!???

(11) Antifa doesn't beat fascists to try to get them to change their mind - they know it's a lost cause to try to change a fascist's mind. They do it to scare them into hiding, to demoralize them, and to stop the spread of their hateful ideology. There is no forcing views on others.

Identity groups are groups of people related by specific characteristics with which they identify. Identity groups include:

- Terms relating to race, gender, ethnicity, nationality, etc.
- Professions, e.g., police officers, teachers, politicians.
- Political parties or groups sharing beliefs, e.g., liberals, Democrats, Republicans, conservatives.
- Political movements, e.g. BLM, ANTIFA, MAGA, QAnon.
- Inhabitants of countries, states, cities, etc.

More specifically:

- References to a single (prominent) member who could serve as representative of an identity group do not constitute references to the identity group; e.g. *Trump* does not reference either conservatives or Republicans.
- References to supporters of a single person are references to identity groups, e.g., *Trumpsters* references Trump supporters, *Obama folks* references Obama supporters.
- References to a corporation or business do not constitute references to identity groups, e.g., *Levi Strauss*, *NFL*, *USTA*, since the members do not necessarily identify with the corporation as a group.
- References to governmental organizations or groups associated with political beliefs do not constitute references to identity groups, e.g., *DNC*, *FBI*, *NRA*, since the members do not necessarily identify with the organization.

Fake identity insults are insults stating that a group of people are not true members of an identity group. These can be via nicknames or in general statements. Fake identity insults are considered *identity abuse*. An example of a tweet with a fake identity insult is shown in example (12), where *RINO* is short for "Republican in Name Only".

(12) @USER Who will abuse it the most: liberals or RINOS?

Name calling are all occurrences of an identity group being referred to by a nickname, often hurtful or derogatory.

3.2 Annotation Guidelines

The annotators developed annotation guidelines before annotating for consistency. There are three types of annotation used in this project, each of which is described below. Relevant definitions are provided above.

1) Type of Abuse Possible labels are: Identity Abuse, Non-Identity Abuse, and Non-Abuse.

Identity Abuse: A tweet contains abusive language towards one or more identity groups.

- Tweets in which the only abuse is targeting a Twitter mention (@USER) are not Identity Abuse.
- Tweets that use identity groups as adjectives in derogatory or insulting ways are Identity Abuse, e.g., "leftist dictatorship", "liberal wagon".
- Tweets targeting a single member of an identity group are not Identity Abuse.
- Tweets that criticize a belief or policy tied to an identity group are not Identity Abuse. However, tweets that criticize those who share a belief or support a policy are Identity Abuse. E.g., "Gun control is stupid": not Identity Abuse; "People who believe gun control are stupid": Identity Abuse.

Non-Identity Abuse: A tweet contains abuse, but said abuse does not target an identity group. (See definition of Abuse and Identity Group.)

Non-Abuse: A tweet does not contain any kind of abuse.

2) Type of Identity Abuse Possible labels are: Explicit and Implicit Mention of Identity Group. This annotation type is only completed if the tweet has been labeled as Identity Abuse.

Explicit Reference: A tweet explicitly names the referenced identity group. That is, the identity group is referenced via its official/common name (see definition of Name Calling), e.g., "liberals", "immigrants" .

Name calling is labeled as explicit reference if the identity group does not need to be inferred and the official or common name of the identity group is present, clearly segmentable, and has typical pronunciation, e.g., "FemiNazis" targets *feminists*, "libtard" targets *liberals*, "dems" targets *Democrats*.

Implicit Reference: A tweet implicitly references the identity group. This can be via Name Calling or by requiring the reader to infer the identity group from contextual information.

Name calling is labeled as implicit reference if the identity group needs to be inferred, or if the official or common name of the identity group is not present, not clearly segmentable, or has atypical pronunciation, e.g., "snowflake" targets *liberals* or *Democrats*, "illegals" targets *immigrants*, "demoncrats" targets *Democrats*.

Implicit reference name calling also includes referencing an identity group with a hyperbolic or exaggerated name. Often, these alternative names target political parties, political movements, organizations, etc.; they are extreme and exaggerate or misrepresent a small subset of the target identity group's characteristics or beliefs. Additionally, name calling can occur via stereotyping, e.g., "Communists" may target *liberals*, "Brown Shirts" targets *ANTIFA*, "blue-haired pronoun-havers" targets *liberals*.

3) Target Identity Group This annotation type is only completed if the tweet has been labeled as Identity Abuse. This annotation field indicates the identity group targeted by a tweet. Consequently there is not a fixed set of labels.

Whenever possible, we copy the identity group directly from the tweet without alteration. We do not interpret identity groups. See examples in (13) and (14).

If labeled as an Implicit Reference, we provide the name of the targeted identity group to the best of our ability.

Identity Groups with Adjectives: Identity that is communicated via adjectives should not be indicated, unless the identity group is the insult itself. See examples in (15) – (17).

Lists of Identity Groups: If identity groups are presented in a list, we indicate all included identity groups. See example in (18).

(13) Liberals are stupid Target: liberals

(14) MAGAs are stupid Target: MAGAs

(15) Christian Conservatives Target: conservatives

(16) socialist, feminist liberals Target: liberals

(17) liberal voters Target: liberals

(18) Liberals, Democrats, and ANTIFA Target: liberals, Democrats, ANTIFA

Category	Label	Count
Type of abuse	identity	395
	non-identity	393
	non-abuse	10
Type of identity abuse	explicit	350
	implicit	25
	N/A	423
Overall		798

Table 1: Statistics of the annotated dataset after conflating the annotations to the majority vote.

3.3 Dataset Statistics

The final dataset comprises 798 tweets. For traditional machine learning purposes, we created a version with single labels by choosing the majority vote among annotators. The distribution in classes for the first two annotation types after the majority vote is shown in Table 1.

4 Inter-Annotator Agreement

Annotators performed the annotations independently, without discussing individual examples (even though general annotation decisions were discussed). This allows us to report inter-annotator

agreement among the five annotators. We report Fleiss' kappa (Fleiss, 1971)[1].

For the identity category, the overall Fleiss' kappa was 0.540 with a 95% confidence interval of [0.521, 0.560]. For the reference category, Fleiss' kappa was 0.591 with a 95% confidence interval of [0.572, 0.611].

Generally, the interpretation of Fleiss' kappa follows the interpretation by Landis and Koch (1977), which is based on the conditions of two annotators and two classes. If we interpret our kappa by those standards, our annotators have moderate agreement, although those standards are likely not the best fit for our conditions. However, given that this is a rather subjective task, we interpret this agreement as indication that our annotation guidelines are sufficient. The inter-annotator agreement shows that the variety in which implicit hate speech occurs makes it difficult to label it as such, even for educated native English speakers (all annotators were undergraduate students at a US university).

An analysis of the posts that have garnered the most disagreement shows that many of them require more context for interpretation. We show examples in (19)–(21).

(19) @USER Great Gun Control! Takes Concentration and Steady Hand! Way to Go Girl! URL

(20) Thank you Father God the American people will begin to see the truth about our government corruption @USER @USER @USER URL

(21) @USER But But But we need more gun control right? Oh wait... GUESS THAT DOESNT STOP CRIMINALS, only law abiding citizens. Which is known. So.... think on why they" want to CONTROL you and disarm you."

In all the above examples, there is no clear majority label among annotators. Each post was annotated as non-abusive by at least one annotator, identity abuse by at least one annotator, and non-identity abuse (but still abusive) by at least one annotator. The presence of "@USER" in each tweet suggests that there likely exists a thread of posts that each tweet is in response to, but without the necessary context of those threads, annotators struggled to determine the intent of the poster and the target of the potentially abusive language.

[1]We used the JASP implementation, https://jasp-stats.org/.

5 Conclusion and Future Work

We have described the annotation process for a corpus of hate speech directed towards identity groups. Throughout our annotation process, we experienced the difficulty and nuance required to define implicit hate. Since implicit hate is not tied to specific words, it is important to use a diverse and complex set of examples, in both the target of the abuse as well as syntactic structure.

The annotated dataset is freely available at https://github.com/donnieparent/X490-hate-speech/tree/main. We have provided two versions, one with all five annotations per example, and one with the a single label, determined by the majority vote.

In future work, we will train models to detect implicit identity abuse, to determine to which degree the variety we have in our dataset is important for identifying identity hate. For a first baseline attempt to determine whether the syntactic bias introduced by Wiegand et al. (2022) influences classificatin results, see (Parent et al., 2025).

Acknowledgments

We are grateful to Annika Shankwitz, who helped with many parts of the project.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France.

J.L. Fleiss. 1971. *Statistical Methods for Rates and Proportions*. John Wiley.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Holly Lopez and Sandra Kübler. 2025. Context in abusive language detection: On the interdependence of context and annotation of user comments. *Discourse, Context & Media*, 63:100848.

Donnie Parent, Nina Georgiades, Charvi Mishra, Khaled Mohammed, and Sandra Kübler. 2025. On the interaction of identity hate classification and data bias. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, CA.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612, Seattle, United States.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Michael Wiegand, Jana Kampfmeier, Elisabeth Eder, and Josef Ruppenhofer. 2023. Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16280–16297, Singapore.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, MN.