Untangling Hate Speech Definitions: A Semantic Componential Analysis Across Cultures and Domains

Katerina Korre[1] and **Arianna Muti**[1] and **Federico Ruggeri**[2] and **Alberto Barrón-Cedeño**[1]
[1]DIT, University of Bologna, Forlì, Italy
[2]DISI, University of Bologna, Bologna, Italy
{aikaterini.korre2, arianna.muti2, federico.ruggeri6, a.barron}@unibo.it

Abstract

Hate speech relies heavily on cultural influences, leading to varying individual interpretations. For that reason, we propose a Semantic Componential Analysis (SCA) framework for a cross-cultural and cross-domain analysis of hate speech definitions. We create the first dataset of hate speech definitions encompassing 493 definitions from more than 100 cultures, drawn from five key domains: online dictionaries, academic research, Wikipedia, legal texts, and online platforms. By decomposing these definitions into semantic components, our analysis reveals significant variation across definitions, yet many domains borrow definitions from one another without taking into account the target culture. We conduct zero-shot model experiments using our proposed dataset, employing three popular open-sourced LLMs to understand the impact of different definitions on hate speech detection. Our findings indicate that LLMs are sensitive to definitions: responses for hate speech detection change according to the complexity of definitions used in the prompt.

Warning: *This paper contains offensive language that might be triggering for some individuals.*

1 Introduction

The infeasibility of formulating a universally accepted definition for hate speech and other related concepts (such as toxic language, cyberbullying, and misogyny) is a much discussed topic that permeates not only Natural Language Processing (NLP) research (Fortuna et al., 2020; Khurana et al., 2022; Pachinger et al., 2023; Korre et al., 2023; Nghiem et al., 2024) but also expands into the legal and social science fields (Maussen and Grillo, 2014; Flick, 2020; Zufall et al., 2022; Guillén-Nieto, 2023). The lack of a clear definition due to cultural diversity hinders the development of models, as it is unclear what criteria they should be trained to detect. For instance, consider two definitions, A and B, where only A covers sexual orientation and political opinion criteria. The statement *"Collectivists are Faggots"* should be labeled as hate speech according to A, and as not hate according to B since B lacks the above-mentioned criteria. Cultural perspectives influence how hate speech is perceived; datasets consist of statements produced by individuals within a culture, so the biases reflect, to some extent, the values, norms, and ethics of that culture (Bagga and Piper, 2020; Hershcovich et al., 2022). Since most NLP research focuses on English-language data (Søgaard, 2022), this cultural dimension is often overlooked, resulting in biases that favor English-speaking cultures.

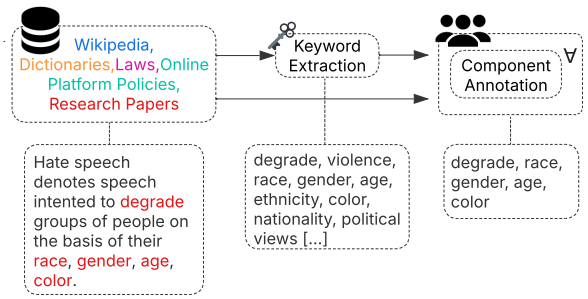

Figure 1: `HateDefCon` creation pipeline.

Current NLP approaches are not adequately equipped to address the cultural dependency of hate speech. Existing monolingual hate speech classifiers often lack cultural awareness (Lee et al., 2024). Prevailing hate speech taxonomies tend to focus more on legal or academic definitions rather than incorporating cultural dimensions, a gap that can prove detrimental, as hate speech per se and hate speech regulation might influence societal discourse, relationships, and cultural norms, potentially shaping how people interact and express themselves (Hietanen and Eddebo, 2023).

Inspired by the compositionality principle (Hinzen et al., 2012), we introduce a component-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3184–3198

annotated resource for hate speech definitions, the HateDefCon dataset. Figure 1 shows our HateDefCon creation pipeline. We propose a Semantic Componential Analysis (SCA), in which we define a hate speech *definitional component* as a fundamental element or criterion used to define what constitutes hate speech in terms of target, intention/purpose, and act/means. We define *definitional hate speech domains* as the contexts where hate speech definitions emerge. This study focuses on five such domains: legislation, Wikipedia, online dictionaries, research papers, and conduct policies from online platforms or technological companies. In addition, we look at the cultural representation of hate speech. We use *culture*, as the term that encompasses language, ideas, beliefs, customs, codes, institutions, tools, techniques, among other elements.[1] Distinguishing between culture and language in hate speech definitions is challenging because languages are not confined to a single culture or country. For example, Arabic is spoken across multiple nations with diverse cultural norms, making it difficult to attribute a hate speech definition to a specific cultural context. Additionally, if definitions are collected in a particular language without information on the cultural or national background, it becomes harder to account for variations in meaning, intent, and societal impact across different regions. Unfortunately, most definitions lack any kind of cultural information. We highlight the need for cross-cultural and cross-domain approaches to define hate speech and argue that such definitions should be context-specific, be it cultural, legal, or academic. We advocate for grounding hate speech definitions within these particular domains. This is in line with best practices for tackling subjective tasks (Rottger et al., 2022), in which the guideline - or the definition - chosen should consider the downstream use, i.e. the context.

In this work, we focus on three research questions: (1) What are the differences among various definitions of hate speech? (2) What is the diversity of these definitions? (3) How do definitions with different components affect the predictions of Large Language Models (LLMs)? Our contributions are both theoretical and practical. On the theoretical front, our cross-cultural and cross-domain analysis of definitions shows significant variation in components, ranging from broad definitions to highly specific ones. Even among the more detailed definitions, which address aspects like the target of hate speech, the intent, and the methods of expressing it, there are differences in their components. On the practical side, we assess whether LLMs respond better to certain definitions, potentially revealing underlying biases. Our results reveal that definitions vary in their components while domains also borrow definitions from one another. When research is culturally specific, borrowing from other domains can be problematic, as it may introduce elements that do not align with the intended cultural context.

The contribution of this paper is threefold:

- We propose a semantically informed framework and use it to create HateDefCon, the first resource for hate speech definitions that amounts to 493 items.
- We conduct a cross-domain and cross-cultural analysis of hate speech definitions and their components.
- We assess whether LLMs exhibit a preference for certain definitions, potentially revealing underlying biases.

2 Related Work

To our knowledge, no previous study has approached the analysis of hate speech definitions from the perspective of semantic componential analysis. However, there have been efforts in NLP that focus on decomposing certain concepts into attributes. We review some of these key approaches, while also discussing how the challenge of defining hate speech has been addressed within the NLP community thus far.

2.1 Decomposition Analysis in NLP

Some approaches that focus on decomposing concepts into further attributes are *concept analysis*, and *conceptual primitives*. For example, Skuce and Meyer (1990) propose using knowledge engineering technology for concept analysis. More preferred in the field of NLP is Formal Concept Analysis (FCA), which focuses on the relationships between sets of objects and their attributes within a formalized mathematical framework, creating concept lattices to represent these relationships (Kamphuis and Sarbo, 1998). FCA is mainly used for ontology construction purposes (Li et al., 2005;

[1] https://www.britannica.com/topic/culture

Moraes and Lima, 2012; Juniarta et al., 2022). Pericliev and Valdes-Perez (1998), on a similar note, perform a concept analysis, involving distinguishing classes based on feature values, which is then applied to linguistic tasks.

At the front of conceptual primitives, i.e. basic units of meaning that cannot be broken down into simpler components within the context of the theory or system they belong to (Smith, 1985), Cambria et al. (2016) introduce a resource utilizing hierarchical clustering and dimensionality reduction. In their follow-up work, Cambria et al. (2018) combine sub-symbolic and symbolic AI to automatically discover conceptual primitives from text and link them to commonsense concepts and named entities in a three-level knowledge representation for sentiment analysis. Similarly, Macbeth (2020) decomposes knowledge and meaning into fundamental perceptual and cognitive structures, improving the interplay between language, expressions, and ontology, and addressing language variation and paraphrasing challenges.

Compared to FCA or conceptual primitives, our SCA framework is a more granular human-driven approach to capturing semantic subtleties at the component level, with particular attention to linguistic context and human interpretability. For example, some domains might emphasize the intent behind hate speech, while others might focus on the effect on the target. With SCA we can pinpoint these differences systematically. In addition, as hate speech is studied across disciplines like law and linguistics, each discipline might have a slightly different understanding of what hate speech is or how it functions. SCA creates a common framework by identifying the shared or different semantic components of hate speech definitions used in these fields, making interdisciplinary research more consistent.

2.2 Hate Speech and the Issue of Definitions

Recent NLP work has focused on comparing harmful language definitions (abusiveness, toxicity, offensiveness, etc.) rather than comparing hate speech definitions per se. Fortuna et al. (2020) focus on clarifying applied categories and homogenizing different datasets by treating class representations as FastText vectors rather than relying on their definitions. While Fortuna et al. (2020) compare terms that are close semantically, such as toxicity, abusiveness, and aggressiveness, our work focuses exclusively on hate speech definitions. Pachinger et al. (2023) review definitions of uncivil, offensive, and toxic comments across 23 papers from various fields, aiming to foster unified scientific resources. Their work highlights the need for consistent terminology across disciplines to enhance clarity and application in scientific research. We embrace this mindset presented in both Fortuna et al. (2020) and Pachinger et al. (2023) by offering a resource that includes a wide range of hate speech definitions —from general to specific, and from various sources— allowing researchers to select from established, culturally relevant options. This helps to avoid the creation of custom definitions when unnecessary, which could introduce bias into experimental models, beginning with the annotation process.

The work that is more similar to ours is the one by Khurana et al. (2022), who develop hate speech criteria with input from legal and social science perspectives to help researchers create precise definitions and annotation guidelines. They propose five criteria: target groups, dominance, perpetrator characteristics, type of negative group reference, and potential consequences/effects. Rather than prescribing a single definition, they offer a meta-prescriptive, modular approach, allowing for adjustments based on specific tasks. This approach emphasizes the role of subjectivity in definitions and addresses the issues arising from varying and vague definitions, which can lead to inconsistencies and problematic expectations about dataset annotations. We build upon the work of Khurana et al. (2022) by introducing a more fine-grained framework, and extend the study by examining the impact of the components when applying different definitions when prompting LLMs.

3 Semantic Componential Analysis

Semantic Componential Analysis (SCA) is a linguistic technique used to break down the meanings of words or phrases into their constituent parts or features. This method, central to structural linguistics, has been in use since the 1950s (Lounsbury, 1956; Goodenough, 1956; Nida, 1975). It is based on the *principle of compositionality*, which states that the meaning of a complex expression is derived from the meanings of its components and the rules governing their combination (Hinzen et al., 2012). SCA involves compiling a detailed list of specific examples for each term within a group

of contrasting terms. Each example is described using a set of relevant attribute dimensions (Kronenfeld, 2005). SCA primarily examines words through organized sets of semantic features, which are marked as 'present' or 'absent', using +/- symbols (Geeraerts, 2006). Appendix A shows an example.

SCA is an approach that enables us not only to break down terms into their individual components, but also to explore various types of meanings as categorized by Leech (1990). By comparing terms side by side, we can enhance our analysis, going beyond the conceptual meaning (the stable meaning across contexts), touching upon the connotative and social meanings. These latter meanings can vary depending on cultural and social contexts.

Kronenfeld (2005) describes the identification of components. One approach involves systematically alternating attributes to determine which distinctions are essential for differentiating terms. Another method is to gather descriptions from informants about differences between terms or subsets of terms, gradually building a set of potential semantic components based on these descriptions. The methodology presented in this paper draws inspiration from the latter approach.

4 The `HateDefCon` Dataset

We analyze and compare hate speech definitions via SCA to identify potential cross-domain and cross-cultural differences. We collect 493 definitions from five different domains in which we can find hate speech definitions (§4.1). We provide information about the respective sources and a final corpus analysis to gain cross-domain and cross-cultural insights (§4.2). Figure 1 summarizes the `HateDefCon` dataset creation pipeline. We collect definitions from the selected domains, we extract keywords which we use to annotate as components for each definition. `HateDefCon`, along with all the results of the experiments is publicly available.[2]

4.1 Data

Hate Speech Laws. We source definitions from the Global Handbook of Hate Speech Laws website,[3] a comprehensive resource that provides access to existing hate speech legislation from countries worldwide, including the United Nations and European Union levels. The legislations are already available in English. We exclude countries for which their legislative texts were not available. The full list of the countries is available in Appendix B.

Wikipedia. We scrape Wikipedia articles in different languages and manually extract the relevant portions of the definitions. We automatically translate extracted portions into English.[4]

Dictionaries. We consult several dictionaries across multiple languages. As with the Wikipedia definitions, we translate the extracted content into English using machine translation.

NLP Research Papers. We use the Hate Speech Dataset Catalogue [5] to navigate through research articles and datasets, focusing specifically on definitions of hate speech. We do not include other related terms like toxicity or abusive language, as our primary focus is hate speech.

Online Conduct Policies. We use the conduct policies from online platforms or technology companies that address the use of hate speech or harmful language online.[6]

4.2 Component Annotation

We engage three annotators for SCA labelling. Annotators are proficient in English, have experience in annotating tasks, and are familiar with hate speech research. All three of them annotate the entire corpus. We provide annotators with detailed instructions on annotating definitions following our framework, as outlined in Appendix C. To compile an initial set of components, we employ keyword extraction and $tf\text{-}idf$ on the collected definitions. A manual review reveals that these automated methods fail to capture some components, highlighting the necessity for human annotation. Despite this limitation, these computational techniques offer a valuable starting point. Using the resulting list, annotators then perform binary annotations, indicating the presence or absence of each component (Kronenfeld, 2005).

[2] https://github.com/katkorre/SCA-of-Hate-Speech.git

[3] https://futurefreespeech.org/global-handbook-on-hate-speech-laws/.

[4] https://www.deepl.com/en/translator.

[5] https://hatespeechdata.com/

[6] The conduct policies were collected from X, Meta, Microsoft, Pinterest, Snapchat, YouTube, Reddit, TikTok, and Discord.

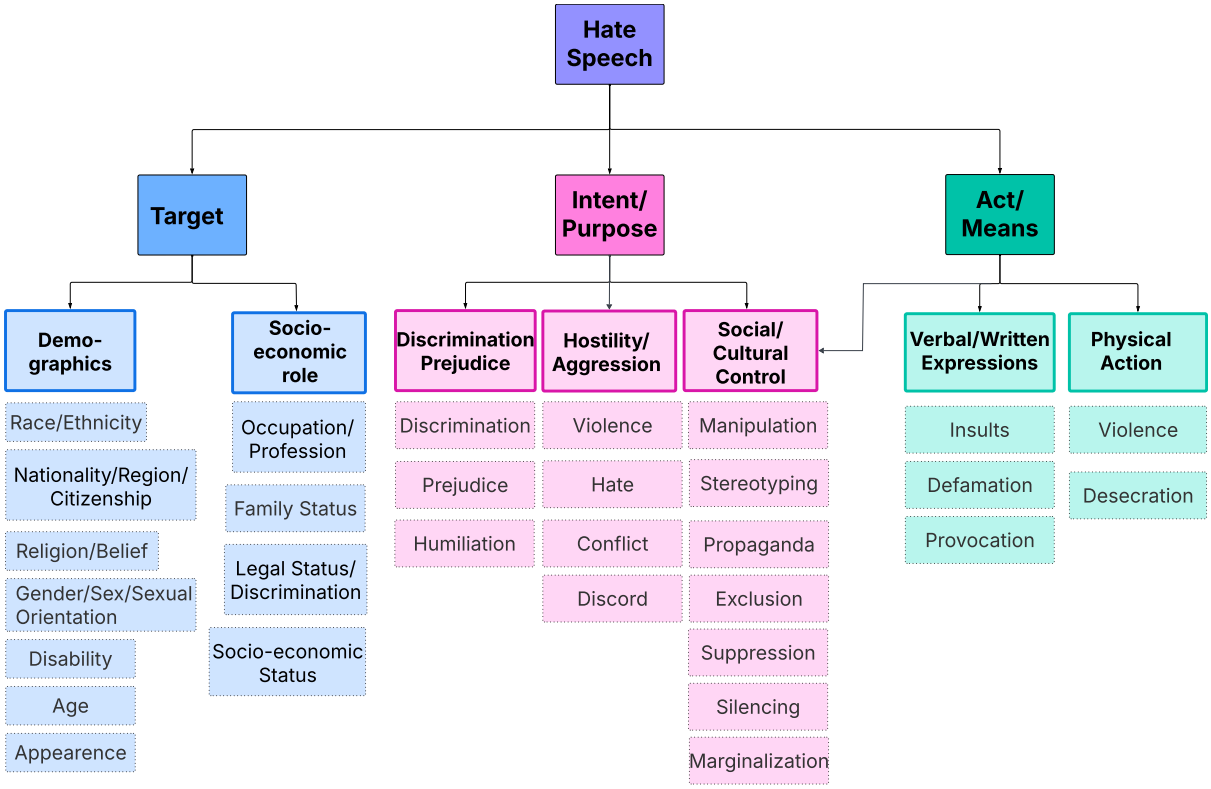

Figure 2: SCA component hierarchy in `HateDefCon`.

To ensure objectivity in the annotation process and a very fine-grained representation of the components, we establish that a definition contains a given component if a derivative of the component's word appears in the annotation. For instance, if the word 'abusive' is present in the definition, the annotator marks the component 'abuse' as present. A challenge we encounter during annotation is the treatment of synonyms. For instance, the words 'harm' and 'hurt' can be considered synonymous in the context of hate speech definitions, but they are not derivatives of the same root word.[7] To maintain consistency and avoid subjective interpretation, which could lead to different grouping of the synonyms by the annotators, we only annotate derivatives of the target term (e.g., marking 'harm' only if words like 'harmful' or 'harmed' appear). Additionally, missing components are labeled as 'undefined component'. This process allows us to comprehensively gather all components deemed crucial by the annotators.

To finalize the dataset, we keep all annotated components, as a manual inspection showed that most of the disagreement derived from the fact that one of the annotators missed a component. The average inter-annotator agreement (IAA) measured via Cohen's kappa (McHugh, 2012) is 0.64 (see Appendix D for more details). Figure 2 reports the component hierarchy resulting from the annotation study (see Appendix E for a fine-grained version of the hierarchy). Figure 3 reports the overall frequency for the top 20 components. The most frequent target components are those mostly related to racism: religion, race, ethnicity and nationality. This indicates a stronger emphasis on aspects of identity that hold deep historical and social significance in discriminatory narratives, compared to biological attributes such as age and sex, which, while still present, appear less prominently.

5 Case Study: Definition Comparison

Qualitative Analysis. We perform a qualitative assessment of the definitions and provide some initial insights into our data. Our observations reveal that many definitional domains, particularly in research papers and Wikipedia articles, frequently borrow definitions from other sources. For instance, 17 out of 30 research papers reference definitions from other academic papers, legislation, or platform policies. Wikipedia, on the other hand,

[7] 'Harm' comes from Middle English 'harm', while 'hurt' comes from Middle English 'hurten'

Domain	No. Definitions	Culture Distribution	Components
Law	116	All cultures appear once	Race, Religion, Hate, Nationality, Ethnicity
Wikipedia	49	All cultures appear once	Religion, Gender, Sexual Orientation, Race, Disability
Research Paper	29	English (13), German (3), Arabic (2), Indonesian (2), the rest appear once	Gender, Religion, Sexual Orientation, Ethnicity, Race
Dictionary	21	English (7), Italian (5), the rest appear once	Attack, Gender, Sexual Orientation, Religion, Hate
Platform	278	All cultures appear once per platform	Disability, Ethnicity, Gender, Nationality, Race
Total	493		

Table 1: Number of definitions, culture distribution, and top 5 components per domain.

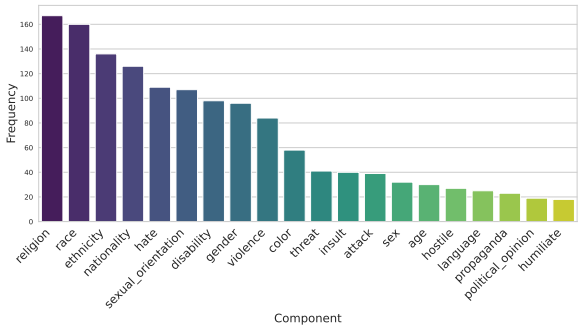

Figure 3: Top 20 most frequent components across all definitions.

often relies on definitions from the Cambridge Dictionary. Through translation, definitions might reflect the culture of the source text, and in collaborative projects, cultural elements from multiple backgrounds may become blended.

Distributions. Table 1 reports culture distribution for each domain in `HateDefCon`. A great disparity is evident with the English definitions, which appear most often in research papers and dictionaries, while platform policies are always the same text but translated in different languages. Definitions from other cultures appear fewer than 5 times, and most occur only once in the dataset. Table 1 also reports the 5 most frequent components per domain. The most common components are related to the target of hate speech, mainly being associated with religion, ethnicity, and gender.

Cross-Cultural Case Study. We employ SCA to describe potential inconsistencies between collected hate speech definitions and the cultural reality through a real case study. Mulki et al. (2019) develop a Levantine Hate Speech and Abusive Twitter dataset in an attempt to bring into the spotlight less-spoken Arabic varieties. However, in their paper, the authors refer to the hate speech definition by Nockleby (2000), which defines hate speech as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic". Levantine Arabic is spoken in many Middle Eastern countries, such as Syria, Jordan, and Israel. Our SCA framework allows us to see if the definitions available for these countries overlap with the one used in Mulki et al. (2019) (Table 2). While they all differ in the intent and act, some targets overlap, although the definition from Mulki et al. (2019) is more comprehensive, considering also sexual orientation and gender, which are not taken into account by the other definitions. Israel's legislation seems to have the most detailed provisions, including concepts of hostility and cultural control, whereas Syria's legislation is narrowly focused on race, and Jordan's legislation emphasizes physical action and terrorism. This variation can affect annotation and prompting, resulting in varied interpretations of hate-related content. Therefore, tailoring prompts to align with specific regional definitions is essential for achieving consistent model behavior for a specific culture.

6 LLM Sensitivity to Definitions

We evaluate the ability of LLMs to perform the task of binary hate speech classification based on the definitions provided. We assess how much LLMs rely on the definitions of hate speech provided rather than their own internal knowledge.

Models and Data. We experiment with three open-source popular state-of-the-art LLMs:

Components	Mulki et al. (2019)	Israel Legislation	Syria Legislation	Jordan Legislation
Target				
Demographics	Ethnicity, Religion, Sexual Orientation, Color, Gender, Nationality, Race	Ethnicity, Religion, Race, Color	Race	Ethnicity, Religion, Race
Intent				
Discrimination	Disparage	Humiliation, Degradation	Race	Prejudice
Hostility	N/A	Hostility, Enmity	N/A	Hurt, Hate, Conflict, Violence
Cultural Control	N/A	Persecution, Degradation	N/A	N/A
Act				
Expression	Disparage	Hostility, Humiliation, Enmity	N/A	Hate
Physical Action	N/A	N/A	N/A	Hurt, Conflict, Violence, Terrorism
Manipulation	N/A	Persecution, Cultural Control	N/A	N/A

Table 2: Comparative analysis of components of Mulki et al. (2019) vs. the legislation of countries where Levantine Arabic is spoken in.

Llama3[8], Mistral2[9], and Phi3-mini.[10] We consider the Gab Hate Corpus (GHC) on online hate speech conversations (Kennedy et al., 2022). GHC consists of 27,665 posts from the social network service gab.com and is annotated for the presence of "hate-based rhetoric" by a minimum of three annotators. The posts are annotated based on a coding typology created by synthesizing definitions of hate speech from legal precedents, existing hate speech coding frameworks, and insights from psychology and sociology. This typology includes hierarchical labels that denote dehumanizing and violent speech, as well as indicators related to targeted groups and rhetorical framing. This dataset captures various dimensions of hate speech, making it well-suited for a detailed and fine-grained analysis and testing with multiple definitions. Although GHC is annotated in a multi-class fashion, we consider a binary setting: hate vs not hate. To limit computational overhead, we select 500 instances from the corpus, equally divided between the two classes.

Setup. We perform our experiments in a zero-shot setting to understand the impact of different definitions (i.e. different components) on model performance without any additional fine-tuning or prompting strategies. To prompt the models, we use three definitions of hate speech. We select the definitions deliberately after manual inspection, aiming to assess whether the varying degrees of component coverage influence the prompting outcomes. This approach includes one definition with medium component coverage, a highly detailed one, and a very general one. We use three definitions: $\texttt{D}_{ghc}$ comes from the GHC dataset, it includes several components but is not the most comprehensive one; $\texttt{D}_{wiki}$ is the definition provided by the Macedonian Wikipedia page, it offers a more detailed and varied perspective; and $\texttt{D}_{dict}$ is a general definition from the Merriam-Webster dictionary (see Appendix F for more details). We set the temperature to zero to exclude variations in the generated responses.

7 Results

Table 3 summarizes the results of the classification task according to the three definitions. Although Llama3 performs consistently better than other models, the complexity of the definitions affects the outputs of the three models differently. In Llama3, the comprehensiveness of the definition is directly proportional to the performance: the more comprehensive the definition is, the more the performance increases; in Mistral2, we observe a reverse tendency. Phi3-mini outperforms Mistral2 in all settings, although being half the size of Mistral2. Additionally, in Phi3-mini, we observe the smallest variety of responses based on the definitions, with $\texttt{D}_{ghc}$ performing the best.

There are cases in which Llama3 refuses to answer for safeguarding the generation process over

[8] `Llama3-8B-instruct.`
[9] `Mistral-7B-Instruct-v0.2.`
[10] `Phi-3-mini-4k-instruct.`

Model	$\mathtt{D}_{ghc}$	$\mathtt{D}_{wiki}$	$\mathtt{D}_{dict}$
Llama3	0.67	0.78	0.70
Mistral2	0.56	0.53	0.58
Phi3-mini	0.60	0.58	0.58

Table 3: Classification performance (F_1score) using difference hate speech definitions.

Def.	$\mathtt{D}_{lev}$	$\mathtt{D}_{isr}$	$\mathtt{D}_{syr}$	$\mathtt{D}_{jor}$
$\mathtt{D}_{lev}$	1.00	0.74	0.59	0.50
$\mathtt{D}_{isr}$	-	1.00	**0.75**	0.67
$\mathtt{D}_{syr}$	-	-	1.00	0.50

Table 4: Agreement across definitions via F_1 score.

True \Predicted	0	1
0	226	24
1	103	147

(a) $\mathtt{D}_{ghc}$

True \Predicted	0	1
0	214	36
1	67	183

(b) $\mathtt{D}_{wiki}$

True \Predicted	0	1
0	222	28
1	102	148

(c) $\mathtt{D}_{dict}$

Figure 4: LLama3 confusion matrices with respect to each hate speech definition. When the model refuses to answer, we set the predicted label to zero.

harmful content. This tendency is also proportional to the completeness of the definition. The less comprehensive the definition, the more the model refuses to answer. For $\mathtt{D}_{wiki}$, the model refuses to answer to 1% of times, 1.6% for $\mathtt{D}_{ghc}$, and 5% for $\mathtt{D}_{dict}$.

Error Analysis. We conduct an error analysis of the answers generated by Llama3. Our error analysis encompasses two aspects: how the model perform considering the definitions against the gold labels, and how the predictions differ according to the definition used. Figure 4 shows the confusion matrices with respect to the three definitions. In all three settings, the false negatives can partly be attributed to the lack of components in the definition. For instance, $\mathtt{D}_{ghc}$ focuses on violence and hate, excluding all other forms of hate speech. As a result, discrimination against a group of people that does not exhibit violence or hate often goes undetected. However, the presence or absence of such components is not a definitive factor. Take the following example sentence: *"Real friends don't let their friends get infected with communism"*. The sentence is incorrectly categorized as non-hateful by $\mathtt{D}_{wiki}$, even though it contains political views as a component. While not conclusive, analyzing the components of misclassified examples can contribute to understanding the reasons behind the classifications.

How do definitions affect the output? We extract all instances for which the answer of Llama3 varies according to the definition, regardless of the ground truth. $\mathtt{D}_{ghc}$ and $\mathtt{D}_{wiki}$ differ in 54 instances, while $\mathtt{D}_{dict}$ differs from $\mathtt{D}_{ghc}$ and $\mathtt{D}_{wiki}$ in 48 instances each. We observe that when the model is prompted with $\mathtt{D}_{wiki}$, it tends to identify more personal attacks, which are often discarded in $\mathtt{D}_{ghc}$ and $\mathtt{D}_{dict}$ as false negatives since such definitions identify group of people as targets of hate speech, rather than individuals. Moreover, with $\mathtt{D}_{wiki}$, LLMs tend to identify more instances that are hateful with respect to political views and undocumented migrants. Indeed, the majority of instances that are correctly identified by $\mathtt{D}_{wiki}$ and misclassified by $\mathtt{D}_{ghc}$ and $\mathtt{D}_{dict}$, contain terms like liberals, communists, and illegal aliens with a negative connotation. In short, we see that the models indeed carry their own biases which are evident in how they classify instances even if we change the definition in the prompt, indicating that they rely more on their own internal knowledge than solely on the definition.

Cross-Cultural Analysis. We go back to the case study of the Levantine dataset (Table 2) to explore how these definitions affect prompting and whether cultural biases may arise as a consequence. We compare predictions for the definition from the original Levantine dataset (Mulki et al., 2019) ($\mathtt{D}_{lev}$), the one for Syria ($\mathtt{D}_{syr}$), the one for Israel ($\mathtt{D}_{isr}$), and the one for Jordan ($\mathtt{D}_{jor}$). The four definitions are in Appendix G. We consider Llama3 for our analysis as our best-performing model. Llama3 achieves the top performance (F_1=0.74) with $\mathtt{D}_{isr}$ and $\mathtt{D}_{syr}$, exceeding most of the results with the original definition. This is an interesting result since $\mathtt{D}_{syr}$ is the least comprehensive definition, with only one target component: *race*. In contrast, using $\mathtt{D}_{lev}$ leads to significantly lower scores (F_1=0.32). Lastly, with $\mathtt{D}_{jor}$ Llama3 achieves F_1=0.67, falling behind best results, yet still largely outperforming its variant using $\mathtt{D}_{lev}$. We also compute the F_1 score across the four definitions. Each time, we assume that one definition represents the gold standard when compared to another one. Table 4 shows the results. The highest F_1 (0.75) is

Def.	D_{isr}		D_{syr}		D_{jor}	
	P	N	P	N	P	N
D_{lev}	55	309	55	295	55	339
D_{isr}	-	-	190	288	157	299
D_{syr}	-	-	-	-	166	294

Table 5: Number of overlapping instances in the cross-cultural setting with separate positive (P) and negative (N) values.

reached between D_{syr} and D_{isr}, followed by D_{lev} and D_{isr}, with a small 0.01 point difference. However, in terms of components, the definitions used in Israel and Jordan are the most similar, as they share three target components: ethnicity, religion, and race. Table 5 shows the number of overlapping instances across the four definitions, divided into positive (P) and negative (N). The agreement between D_{syr} and D_{isr} is confirmed by the number of overlapping instances, as they have the greatest agreement with 478 instances. In all cases, most of the overlapping instances are on the negative class, as it is the class that the model tends to overpredict.

8 Conclusions

We introduced HateDefCon, a comprehensive hate speech definition dataset. HateDefCon provides detailed component annotations, capturing the target, intent/purpose, and act/means of collected hate speech instances that allow comparing hate speech definitions. Our analysis of HateDefCon, revealed a lack of cultural diversity in existing definitions. This is primarily because only legislative sources referred to specific cultures. Wikipedia definitions are also often the result of collective contributions or translations of the original English text, making them less culturally distinct. There is also a variation in terms of components, especially when one language can refer to multiple cultures, such as in the case of Levantine Arabic. Moreover, our experiments showed that LLMs are sensitive to the employed hate speech definitions, where, in some cases, more comprehensive definitions lead to better results.

Our work underscores two key considerations for hate speech detection research: (a) Definitions to be incorporated into the annotation guidelines or prompts must be specific to the task and should clarify the level of comprehensiveness or generality required for that task. This consideration should also align with model selection, as some models perform better with general definitions while others with more comprehensive ones. (b) Definitions must be relevant to the language and the target culture. This may involve referring to hate speech legislation, to understand what constitutes hate speech in the given culture, otherwise clarify that no culture is considered.

In future work, we plan to develop a community knowledge base of hate speech definitions to foster research in several emerging cultures and domains in this field.

Limitations

Despite our efforts to gather as many definitions as possible for HateDefCon, finding non-English or culture-specific definitions remains a challenge. As discussed in Section 4, most of the non-English definitions we obtain are sourced from legislation, while in the case of online platforms and tech companies, they are often translations of the English entries. This limits the breadth of cross-cultural analysis and restricts the scope of our study.

Acknowledgments

K. Korre's research is carried out under the project *RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete*, in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118. F. Ruggeri is partially supported by the project European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso, PE00000013 - "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI". Arianna Muti's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE).

Ethics Statement

We clarify that, while some definitions may be more comprehensive or specific than others and acknowledge that definitions can be culture-bound, no definition should be considered inherently superior to another based solely on cultural context. As evident from the literature, definitions often reflect the values of the culture they originate from. The absence of certain components in a definition (such as not including gender as a target) may reflect a more conservative stance in some cases. However,

the purpose of this work is not to scrutinize cultures by using hate speech definitions as a proxy, but rather to highlight the fundamental differences that must be considered when applying these definitions for hate speech detection. In other words, the definitions should not only be tailored to the task at hand, but also be specific to the cultural context they represent.

With regard to the annotation, all annotators gave their full consent to participate in this study.

Working setting

Our experiments were carried out on an Nvidia A100 GPU, running for 20 minutes on average for each experiment. To limit computational overhead, we did not experiment with more powerful models.

AI-assisted technologies

ChatGPT 3.5 has been used to improve the language and readability of the paper. We take full responsibility for the content, which has been properly reviewed and edited to reflect our own methods.

References

Sunyam Bagga and Andrew Piper. 2020. Measuring the effects of bias in training data for literary classification. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 74–84, Online. International Committee on Computational Linguistics.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Bjoern Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, Osaka, Japan. The COLING 2016 Organizing Committee.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Caterina Flick. 2020. The legal framework on hate speech and the internet good practices to prevent and counter the spread of illegal hate speech online. In *Language, Gender and Hate Speech A Multidisciplinary Approach*. Fondazione Università Ca' Foscari.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Dirk Geeraerts. 2006. Componential analysis. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 709–712. Elsevier, Oxford.

Ward H. Goodenough. 1956. Componential analysis and the study of meaning. *Language*, 32(1):195–216.

Victoria Guillén-Nieto. 2023. *Hate Speech*. De Gruyter Mouton, Berlin, Boston.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Mika Hietanen and Johan Eddebo. 2023. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458.

Wolfram Hinzen, Edouard Machery, and Markus Werning. 2012. *The Oxford Handbook of Compositionality*. Oxford University Press.

Nyoman Juniarta, Olivier Bonami, Nabil Hathout, Fiammetta Namer, and Yannick Toussaint. 2022. Organizing and improving a database of French word formation using formal concept analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3969–3976, Marseille, France. European Language Resources Association.

Vera Kamphuis and Janos Sarbo. 1998. Natural language concept analysis. In *New Methods in Language Processing and Computational Natural Language Learning*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-cedeño. 2023. Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230, Toronto, Canada. Association for Computational Linguistics.

David B. Kronenfeld. 2005. Cognitive research methods. In Kimberly Kempf-Leonard, editor, *Encyclopedia of Social Measurement*, pages 361–374. Elsevier, New York.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Geoffrey N. Leech. 1990. *Semantics: The Study of Meaning*. A Penguin book. Penguin Books.

Sujian Li, Qin Lu, and Wenjie Li. 2005. Experiments of ontology construction with formal concept analysis. In *Proceedings of OntoLex 2005 - Ontologies and Lexical Resources*.

Floyd G. Lounsbury. 1956. A semantic analysis of the pawnee kinship usage. *Language*, 32(1):158–194.

Jamie C. Macbeth. 2020. Enhancing learning with primitive-decomposed cognitive representations. In *Proceedings of Machine Learning Research*, volume 131 of *IWSSL*, pages 98–107, Northampton, MA, USA. Smith College, Smith College.

Marcel Maussen and Ralph Grillo. 2014. Regulation of speech in multicultural societies: Introduction. *Journal of Ethnic and Migration Studies*, 40(2):174–193.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Sílvia Moraes and Vera Lima. 2012. Combining formal concept analysis and semantic information for building ontological structures from texts : an exploratory study. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3653–3660, Istanbul, Turkey. European Language Resources Association (ELRA).

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Huy Nghiem, Umang Gupta, and Fred Morstatter. 2024. "define your terms" : Enhancing efficient offensive speech classification with definition. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1309, St. Julian's, Malta. Association for Computational Linguistics.

Eugène Albert Nida. 1975. *Componential Analysis of Meaning: An Introduction to Semantic Structures*. Mouton, The Hague.

John T. Nockleby. 2000. Hate speech. In Leonard W. Levy and Kenneth L. Karst et al., editors, *Encyclopedia of the American Constitution*, 2nd edition, volume 1. Macmillan, New York.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

Vladimir Pericliev and Raul E. Valdes-Perez. 1998. A procedure for multi-class discrimination and some linguistic applications. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1034–1040, Montreal, Quebec, Canada. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Douglas Skuce and Ingrid Meyer. 1990. Concept analysis and terminology: A knowledge-based approach to documentation. In *COLING 1990 Volume 1: Papers presented to the 13th International Conference on Computational Linguistics*.

Raoul N. Smith. 1985. Conceptual primitives in the english lexicon. *Paper in Linguistics*, 18(1):99–137.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A legal approach to hate speech – operationalizing the EU's legal framework against the expression of hatred as an NLP task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Semantic Componential Analysis

Table 6 illustrates an example of SCA application.

	Parent	Sibling	Male	Female
Father	+	-	+	-
Mother	+	-	-	+
Brother	-	+	+	-
Sister	-	+	-	+

Table 6: Kinship terms characterized by attribute dimensions. The attribute dimensions include Parent, Sibling, Male, Female. Each example is marked with '+' if the feature is present, '-' if absent.

B Laws from the Global Handbook on Hate Speech Laws

We provide the full list of countries included and excluded from `HateDefCon`.

Included. Afghanistan, Albania, Algeria, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei Darussalam, Bulgaria, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Croatia, Cuba, Cyprus, Czech Republic, Democratic Republic of Congo, Denmark, Djibouti, Estonia, Ethiopia, Fiji, Finland, France, Gabon, Georgia, Germany, Ghana, Greece, Guinea, Guinea-Bissau, Guyana, Haiti, Hungary, Iceland, India, Indonesia, Iran (Islamic Republic of), Iraq, Ireland, Italy, Japan, Jordan, Kenya, Kyrgystan, Latvia, Luxembourg, Myanmar, Malaysia, Malta, Mexico, Moldova, Monaco, Myanmar, Nepal, Netherlands, New Zealand, Norway, Oman, Pakistan, Sweden, Syrian Arab Republic, Tanzania, Timor-Leste, Trinidad and Tobago, Tunisia, Turkmenistan, Uganda, United Arab Emirates, United States of America, Uzbekistan, Venezuela, Zambia, Poland, Portugal, Romania, Russian Federation, Rwanda, Senegal, Serbia, Sierra Leone, Singapore, Slovakia, Somalia, South Africa, South Sudan, Spain, Sri Lanka, Switzerland, Tajikistan, Togo, Turkey, Ukraine, United Kingdom, Uruguay, Vietnam, Zimbabwe.

Excluded. Madagascar, Malawi, Maldives, Mali, Marshall Islands, Mauritania, Mauritius, Micronesia (Federated States of), Mongolia, Montenegro, Morocco, Mozambique, Namibia, Nauru, Nicaragua, Niger, Nigeria, North Macedonia, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Qatar, Republic of Korea, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa, San Marino, Sao Tome and Principe, Saudi Arabia, Seychelles, Slovenia, Solomon Islands, Suriname, Thailand, Tonga, Tuvalu, Vanuatu, Yemen.

Read and Understand the Definition. Carefully read the provided hate speech definition from the source material. Ensure you understand the context and content before proceeding with annotations.
Identify and Annotate Core Components. After reading the definition, review all available components listed in each Excel column. Annotate each component with 1 if the component exists in the definition. Annotate with 0 if the component does not exist in the definition. If a definition is very general, annotate with 1 on the column "General". If you see words/phrases like sexism, or sexist behavior, please annotate by putting one in gender.
Undefined Components. If you detect a component that has not been included in the columns, please use the column "Undefined Component" to write it down.
Add Comments. Feel free to add any relevant comments in the comments column.
Note: If a component or a morphological derivative of the component appears in the definition, mark the respective column with a positive annotation. For instance, if the predefined component is "abuse" and the component "abusive" is found in the definition, place a positive annotation (i.e., 1) in the "abuse" column.

Table 7: `HateDefCon` annotation guidelines.

C Annotation Guidelines

Table 7 reports annotation guidelines for building `HateDefCon` via SCA.

D Inter-annotator Agreement for Componential Annotation

Table 8 reports pairwise IAA on `HateDefCon`.

Annotator Pair	Average Kappa
Annotator_1 vs. Annotator_2	0.75
Annotator_1 vs. Annotator_3	0.64
Annotator_2 vs. Annotator_3	0.64

Table 8: Average Cohen's Kappa scores per annotator pair

E Complete Component Hierarchy and Further Statistics

Table 9 reports the fine-grained SCA hierarchy extracted from `HateDefCon`.

F Hate Speech Definitions for GHC

Table 10 reports the definitions used in our experiments with LLMs on GHC.

G Definitions for Cross-Cultural Analysis

We select three definitions from our corpus. Table 11 shows the prompt and the definitions used for the experiments.

Hate Speech Framework	Categories
Target	**Demographics and Identity** Race/Ethnicity: Race, Ethnicity, Tribe, Color, Nationality, Regional Origin, Genetic Origin Nationality/Region: Nationality, Region, Place of Birth, Place of Origin, Place of Residence, Immigration Status Religion/Belief: Religion, Belief, Creed, Ideology, Philosophical Opinion, Philosophical Ideology, Worldview Gender/Sexual Orientation: Gender, Sexual Orientation, Gender Identity, Sex Change Disability: Disability, Physical Condition, Mental Capacity, Health Characteristics, Ability Age/Appearance: Age, Appearance, Generation
	Social and Economic Roles Occupation/Profession: Occupation, Profession, Job, Employment, Trade Union Membership, Calling Family Status: Family Status, Marital Status, Familial Status, Pregnancy Citizenship/Legal Status: Citizenship, Legal Status, Immigration, Veteran Status, Refugees, Nationality
	Social and Economic Class Socioeconomic Status: Economic Status, Social Class, Financial Status, Wealth, Poverty, Social Origin, Social Strata, Economic/Social Origin Caste/Tribe: Caste, Tribal Affiliation, Ancestry, Descent
Intent/Purpose	**Discrimination and Prejudice** Discrimination: Discrimination, Discriminatory Practices, Exclusion, Marginalization, Segregation, Denigration Prejudice: Prejudice, Bias, Stereotyping, Xenophobia, Ethnocentrism, Bigotry, Contempt, Superiority Humiliation: Humiliation, Demeaning, Belittling, Degrading, Ridiculing, Mocking, Stigmatizing, Inferiorizing, Dehumanizing
	Hostility and Aggression Violence: Physical Violence, Aggression, Abuse, Threat, Brutalization, Persecution, Terrorism, Hostility, Rancor Hate: Hatred, Ill-Will, Animosity, Abhorrence, Detestation, Malice, Anti-Semitism, Racism, Ethnocentricism Conflict: Conflict, Discord, Dissension, Sectarianism, Division, Social Unrest, Civil Unrest, War
	Social and Cultural Control Cultural Control: Cultural Manipulation, Propaganda, Ideological Imposition, Social Control, Social Hatred, Supremacy, Perpetuation of Norms, Customs, Traditions, Cultural Values Exclusion: Exclusion, Social Marginalization, Social Exclusion, Economic Exclusion, Stigmatization, Alienation, Isolation Suppression: Suppression, Silencing, Censorship, Restriction, Limitation of Rights, Deprivation, Harassment
Act/Means	**Verbal and Written Expressions** Insults: Insults, Pejoratives, Slurs, Offensive Language, Derogatory Language, Humiliation, Threat Defamation: Defamation, Slander, Vilification, Disparagement, Discrediting, Ridicule, Mockery Provocation: Provocation, Incitement, Inflammatory Speech, Sedition, Antagonism, Triggering, Threats Misinformation: Misinformation, Disinformation, Propaganda, Deception, Promoting Xenophobia, Racism, and Bigotry
	Physical Actions Violence: Physical Harm, Assault, Attack, Damage to Property, Brutalization, Persecution Exclusion: Exclusion, Segregation, Denial of Rights, Obstructing Rights, Deprivation, Harassment Cultural Actions: Desecration, Desecration of Symbols, National Flag Desecration, Denial of Cultural Identity
	Social and Cultural Manipulation Social Control: Manipulation of Social Norms, Cultural Domination, Supremacy, Cultural Stereotyping, Perpetuation of Prejudice Cultural Exclusion: Cultural Alienation, Cultural Stigmatization, Exclusion from Cultural Activities, Denial of Cultural Identity Economic Suppression: Economic Marginalization, Social Segregation, Restriction of Economic Opportunities, Denial of Economic Rights

Table 9: Hate Speech Framework Hierarchical Structure.

No.	Definition
$\mathtt{D}_{ghc}$	Language that intends to — through rhetorical devices and contextual references — attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred.
$\mathtt{D}_{wiki}$	Hate speech — a term that denotes speech intended to degrade, disturb, or cause violence or actions based on prejudice against persons or groups of people on the basis of their race, gender, age, ethnicity, nationality, religion, sexual orientation, gender identity, disability, language ability, moral or political views, socioeconomic class, occupation or appearance (such as height, weight, and hair color), mental capacity, and any other characteristic. The term refers to both written and oral communication, as well as some forms of behavior in a public place. Hate speech operates outside the law, speech that offends a particular person or group in terms of discrimination against that person or group. According to the law, hate speech is any speech, gesture or behavior, written text, or display that is prohibited because it is likely to incite violence or prejudice against or by a protected individual or group, or that degrades or intimidates a particular individual or group. The law may recognize the protected individual or group according to certain characteristics.
$\mathtt{D}_{dict}$	Speech expressing hatred of a particular group of people.

Table 10: Definitions of Hate Speech. D_{ghc} is the one used to annotate GHC, D_{wiki} is from Wikipedia, and D_{dict} is from the Mariam Webster Dictionary.

Prompt
Read carefully the definition of 'hate speech' provided. Your task is to classify the input text as containing hate speech or not. You can only rely on the definition provided. Respond only with YES or NO. Definition: {`definition`} Text: {`text`} Answer:
Definitions
$\mathtt{D}_{lev}$. Hate speech (HS) is formally defined as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" (Mulki et al., 2019)
$\mathtt{D}_{isr}$. In this Article: "racism" – persecution, humiliation, degradation, a display of enmity, hostility or violence, or causing violence against a public or parts of the population, all because of their colour, racial affiliation or national ethnic origin. Publication of racist incitement is prohibited Article 144B: (a) If a person publishes anything in order to incite to racism, then he is liable to five years imprisonment. (b) For the purposes of this section, it does not matter whether the publication did cause racism, and whether or not it is true. Article 144C: Permissible publication (a) Publication of a true and fair report of an act said in section 144B shall not be deemed an offense under that section, on condition that it was not intended to cause racism. (b) Publication of quotes from religious scriptures or prayer books or the observance of a religious ritual shall not be deemed an offence under section 144B, on condition that it was not intended to cause racism. Article 144D: Possession of racist publication If a person holds a publication prohibited under section 144B for distribution, in order to cause racism, then he is liable to one-year imprisonment, and the publication shall be confiscated.
$\mathtt{D}_{syr}$. Criminal Code Article 306: Any act, piece of writing or speech that is intended to or results in stirring sectarian or racial strife or inciting conflict between sects and the various elements of the nation shall be punished by imprisonment of six months to two years and a fine of one hundred to two hundred Syrian pounds, as well as with prohibition from exercising the rights mentioned in the second and fourth paragraphs of Article 65. Article 65: Every person sentenced to imprisonment or house arrest in misdemeanor cases is deprived throughout the execution of his sentence from exercising the following civil rights: A:The right to assume public employment and services. B: The right to assume jobs and services in managing the affairs of the civil sect or managing the union to which he belongs. C: The right to be a voter or elected in all state councils. D: The right to be a voter or elected in all sects and trade union organizations. E: The right to wear Syrian or foreign medals.
$\mathtt{D}_{jor}$. Criminal Code Section 5: Crimes Harming National Unity and the Coexistence between the Nation's Elements Article 150: Any writing or speech aims at or results in stirring sectarian or racial prejudices or the incitement of conflict between different sects or the nation's elements, such act shall be punished by imprisonment for no less than six months and no more than three years and a fine not to exceed five hundred dinars (JD500). Audiovisual Media Law Article 20(l)(2) prohibits licensed broadcasters from broadcasting hateful, terrorist, violent or seditious material or from promoting religious, sectarian or ethnic strife.

Table 11: Prompt and definitions used in the cross-cultural experiments.