Hertie School

"How would you define hate speech?" - An Analysis on How Formal Education Shapes Individual Conceptions of Hate Speech on Social Media

Master Thesis

by Johanna Mehler

Supervisor: Prof. Dr. Simon Munzert

Master of Data Science for Public Policy

May 2024

Word Count: 7489

Executive Summary

Amidst the concern over hate speech online and its amplification via social media, the discourse on the trade-off between freedom of speech and protection of vulnerable societal groups is receiving great attention. But how inclusive is the discussion? Participation in political discourse is also a question of education. To investigate this within the phenomenon of hate speech on social media, the following research question is addressed: How does formal education shape conceptions of hate speech?

Literature suggests higher education correlates with sophisticated yet ideologically defined political knowledge. It is therefore hypothesized that an academic level of education leads to more sophisticated definitions of hate speech and that academic respondents' definitions are more ideologically defined than those of non-academics. The analysis draws on data from over 19,000 participants across eleven countries, examining hate speech moderation preferences. Using open-text responses, several indicators are created to understand how individuals define hate speech in content and form. Statistical tests compare academic status groups, while regression models explore the impact of academic status, also in interaction with political ideology and hate speech experience. Findings show that the definitions of academics are generally not longer or remarkably more readable. However, they tend to define hate speech more intent-based or narrowly harms-based than non-academics. Academics with higher political interest and left-leaning ideologies provide longer and more readable definitions as those with lower interest and further to the right, while there is no clear pattern for non-academics in either context. This supports the assumption that political attitudes play a greater role in defining hate speech for academics than for non-academics.

Contents

1 Introduction

"Hate speech is one of the most worrying forms of racism and discrimination prevailing across Europe and amplified by the Internet and social media. Hate speech online is the visible tip of the iceberg of intolerance and ethnocentrism."

(Keen et al., 2020, p. 215)

Hate speech, amplified by social media, is increasingly being described as a key threat to democratic societies (Izquierdo Montero et al., 2022; Solovev & Pröllochs, 2022). This is particularly evident in the link between hate speech and a global increase in violence against minorities (Council on Foreign Relations, 2019).

Hate speech has become a daily used tool of nationalist and xenophobic parties in various European parliaments and even governments (Izquierdo Montero et al., 2022), but the political sphere is also involved in combating the spread of hate speech. More and more countries are trying to regulate hate speech and are therefore attempting to include the phenomenon in their criminal codes (Council on Foreign Relations, 2019; Sellars, 2016), taking different approaches. While the United States grant social media companies far-reaching powers to manage their content, countries such as Germany can force companies to remove posts (Council on Foreign Relations, 2019).

These differences are partly due to the lack of a universally accepted definition of hate speech, which leads to difficulties in identifying hate speech in the first place. (Guillén-Nieto, 2023; Munzert et al., 2023). Naturally, this poses a great challenge to the development and coherent implementation of legitimate regulation. Or as Sellars (2016, p. 32) puts it: "A person who says they have an easy solution to the problem of hate speech, or even how to observe and document hate speech, is simply not thinking hard enough".

The United Nations (2019) claim that in addition to governments, also civil society, the private sector and individuals are responsible for tackling hate speech.

Looking at the level of individuals, Munzert et al. (2023) highlight that the perceived harm of speech varies from person to person, and knowledge of the concept may influence this perception. The authors also raise questions about the influence of norms, both in society and educational institutions. At the same time, little is known about individuals' knowledge of hate speech. In one of the few existing studies, 116 students were asked about their concept of hate speech (Jubany & Roiha, 2016). The majority of those surveyed were not at all familiar with the term, but had an intuitive idea of what it meant. When defined for example as "the opposite of love", specific motives for hatred such as ethnicity or religion are not recognized. According to Jubany and Roiha (2016), this prevents a clear differentiation from phenomena such as cyber-bullying and implies the need of pedagogical measures. On the other hand, untrained adults in Munzert et al. (2023) appeared capable of judging hate speech, prompting inquiries into the necessity of specialized education on the topic.

Meanwhile, the call for more education as a tool to prevent and combat hate speech at an individual level is very popular (Estellés & Castellví, 2020; Jubany & Roiha, 2016; Kansok-Dusche et al., 2023; Keen et al., 2020; United Nations, 2019). In their meta-analysis of 266 studies on hate speech published between 2001 and 2020, Izquierdo Montero et al. (2022) find that most studies on hate speech come from the legal sphere (40 percent), discourse analysis (20 percent), and machine learning (12 percent). Besides computational linguistics and technical strategies to combat hate speech, research on hate speech prevention primarily emerges in the domain of media and information literacy (Izquierdo Montero et al., 2022). Still, Izquierdo Montero et al. (2022) attest to a lack of research from the field of communication and education that could provide critical pedagogical perspectives.

To better understand how formal education affects people's concepts of hate speech, this study aims to explore the influence of academic education on several characteristics of people's hate speech definitions. Thereby, the following research

question is addressed: How does formal education shape people's concepts of hate speech?

The study aims to contribute to the knowledge of whether the discussion on hate speech has reached the breadth of society, or whether the trade-off between freedom of expression and the protection of vulnerable social groups is purely an ivory tower discourse in the academic world. Ideally, this study helps future research to determine the usefulness, the right channel and time for educational interventions.[1]

2 Literature and Theory

The context of the research question includes effects of educational attainment on political knowledge and opinion, as well as studies on definitions and attitudes towards hate speech. In addition, relevant methodological papers on the treatment of open-text responses are referenced in 4.1.

2.1 Education and Political Attitudes

If knowledge about hate speech is considered political knowledge, existing research would support the assumption that educational attainment does play an important role in what people know about the concept, given that education is one of the fundamental forces shaping people's political knowledge (Hall, 2018; Weakliem, 2002; Weinschenk & Dawes, 2019).

Fittingly, from an institutional perspective, education offers a popular starting point for taking action against unwelcome social phenomena and promoting democratic coexistence: The work of the Council of Europe against hate speech, for example, relies primarily on educational institutions to carry out human rights education and education for democratic citizenship to initiate the "lifelong learning

[1] All code and supplementary material are available at https://github.com/j-mehler/Hertie-Thesis-Mehler

process of practising democracy" (Keen et al., 2020). Next to technology, education is mentioned by the United Nations (2019) as a key tool for addressing and countering hate speech.

However, existing research is divided as to whether education is an important factor for political opinion (Bobo & Licari, 1989; J. Chan, 2019; Heijden & Verkuyten, 2020) and political interest (Highton, 2009; Witschge et al., 2019).

For certain political attitudes, the influence of formal and especially academic education is a consistent finding. This is reflected in the frequently demonstrated liberalizing effect of formal education on attitudes toward refugees and immigration (Bobo & Licari, 1989; Ceobanu & Escandell, 2010; Heijden & Verkuyten, 2020). Bobo and Licari (1989) attribute a substantial fraction of this education effect to cognitive sophistication, providing a theoretical bridge to individually available knowledge about hate speech as a phenomenon.

However, Highton (2009) did not find a significant effect of graduating from college on political awareness. The author finds that "differences in political sophistication evident after people attend college are already in place before anyone sets foot in a college classroom" (Highton, 2009). Research from Witschge et al. (2019) shows in a differentiated way that although a change in the educational track has only minor effects on voting behavior and trust in institutions, students making transitions in academic education develop a higher level of political interest.

Weakliem (2002) also reports uncertainty about the scope and interpretation of the liberalizing effect. According to his findings, a connection between education and individualist values can be observed, but at the same time education is also associated with a lower level of trust in most institutions. This obviously undermines the belief of many that education is the strongest force for strengthening people's democratic standing.

Since no clear boundaries are drawn in the literature between political knowledge and political opinion, it is not easy to compare the existing evidence. The situation

is complicated: Heijden and Verkuyten (2020) found that endorsement of social conformity and acceptance of group-based inequality were more strongly anchored in political orientation among participants with higher levels of education than among participants with lower levels of education. In particular, education (and thus presumably also political knowledge) and, on the other hand, political orientation were found to be two independent correlates of anti-immigrant and anti-refugee attitudes. This is supported by J. M. Chan et al. (2002): "Past studies in the West have shown that education increases people's support for abstract democratic principles, but not necessarily for concrete policies implementing these principles".

To put it bluntly: What you know and what you want can differ. This is probably just as true in the case of political knowledge and political attitudes. Nevertheless, Heijden and Verkuyten (2020) show that education not only influences the approval of ideological political beliefs, but also contributes to people linking their political orientation more closely to their personal values: "Political sophistication helps people to identify policies that suit their personality" (Heijden & Verkuyten, 2020). The literature reviewed implies that it is needed to repeatedly examine the educational effect on political knowledge and attitudes on a topic- and context-specific basis.

2.2 Defining Hate Speech

"Some of them express or advocate views but do not call for action. Some are abusive or insulting but not threatening. Some express dislike of a group but not hatred, and some of those that do are so subtle as not to be obviously abusive or insulting. Some take a demeaning or denigrating view of a group but wish it no harm and even take a [patronizingly] indulgent attitude toward it."

(Parekh (2012) on the diversity of hate speech)

Neither in Academia, nor in the legal sphere there is a common sense on how hate speech should be defined (Guillén-Nieto, 2023; Kansok-Dusche et al., 2023; Sellars, 2016; United Nations, 2019). In addition, research showed some overlap of

online hate speech with other concepts such as cyberbullying (Kansok-Dusche et al., 2023) and, as Guillén-Nieto (2023) points out, the social phenomenon and the legal concept of hate speech are inherently linked.

So, defining hate speech is difficult, and according to Sellars (2016), "for good reasons given the competing interests at play". This indicates that there is usually a motivation behind the definition of hate speech, which is derived from the respective perspectives of researchers, lawyers, or online platforms. The same seems to apply to political and civic reports (Kansok-Dusche et al., 2023), and legal definitions that originate from different legal cultures (Guillén-Nieto, 2023).

Out of many, two legal theories should be highlighted within the scope of this work: the United Nations' internationally valid classification criteria for hate speech (2012) and the three distinct categories for hate speech definitions developed by Marwick and Miller (2014). In addition, a meta-study by Kansok-Dusche et al. (2023) provides a point of reference for measuring the scope of definitions.

2.2.1 Legal Criteria for Hate Speech

Sellars (2016) and Guillén-Nieto (2023) both list "the Rabat Plan of Action" (United Nations, 2012) as the most successful effort to date for an internationally valid classification of hate speech. The document, which was developed in a multi-stakeholder process, attempted to harmonize the right to freedom of opinion and expression in Article 19 of the International Covenant on Civil and Political Rights (ICCPR, United Nations, 1966) with the prohibition of "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence" (Article 20, ICCPR). It contains six conditions for assessing when speech is severe enough to warrant punishment under Article 20 of the ICCPR, taking into account the diversity of the phenomenon (based on Guillén-Nieto (2023)):

1. The socio-political context

2. The speaker's power compared to the target group

3. The sender's intent

4. Content and form (e.g. style, argumentation)

5. The extent of the speech act (e.g. in a restricted environment or widely accessible to the general public)

6. Likelihood that the speech act incites hatred, hostility or violence against the target group

2.2.2 Approaches towards Hate Speech Definitions

Definitions in any context rarely capture all of these criteria, presumably also to avoid too narrow a scope, which would exclude many cases. At the same time, Marwick and Miller (2014) identified three approaches that scholars use to define hate speech (description taken from Guillén-Nieto 2023, p. 3):

1. **Content-based**: words, expressions, or symbols that are "generally considered offensive to a particular group of people and objectively offensive to society"

2. **Intent-based**: "requires the speaker's communicative intention to incite hatred or violence against a particular minority, member of a minority, or person associated with a minority without communicating any legitimate message"

3. **Harms-based**: "speech that causes the victim harm, such as loss of self-esteem, physical and mental stress, social and economic subordination and effective exclusion from mainstream society"

As a broad definition, the authors themselves use an intent-based approach by suggesting that hate speech is "[. . .] speech that carries no meaning other than hatred towards a particular minority, typically a historically disadvantaged minority" (Marwick & Miller, 2014, p. 17). Kansok-Dusche et al. (2023) also describe

intent as a particularly crucial component in ensuring that hate speech does not become a catch-all term for offensive expressions, but that it fulfills its potential to identify those hateful statements that seriously undermine the peaceful coexistence in a pluralistic society.

Deviating from the categorization by Marwick and Miller (2014), they used the description "potential harm" instead of "harm" in their study, based on Sellar's 2016 objection that the effect of speech and how it is understood by audience is not necessarily obvious.

2.2.3 Scope of Hate Speech Definitions

As a third essential categorization of hate speech definitions, the description of the scope according to Kansok-Dusche is worth pointing out here. In their meta-study on hate speech definitions of adolescents and children, they used a text-based detection approach to screen 18 studies for general terms and descriptive features of hate speech definitions. In this way, a typology of definitions was created based on the total number of descriptive features. In this way, they categorized the hate speech definitions as broad (1-3 descriptive terms), medium (4-5 descriptive terms), or narrow (6+ descriptive terms).

2.3 Hypotheses

The insights from the described literature in 2.1 motivate the inquiry into whether higher education contributes to a more nuanced understanding of hate speech and lead to Hypothesis 1:

H1: An academic level of education leads to a longer and better readable personal definition of hate speech than lower educational levels.

As noted in 2.2, in addition to the knowledge base of a definition, motivation or political interest are other important factors that play a role in defining hate speech. The described alignment of political knowledge and personal values among people

with a higher level of education in 2.1 motivates Hypothesis 2:

H2: Academic respondents' definitions of hate speech are more politically defined than those of non-academics.

2.4 Variable Setup

To identify the effect of academic education on the form and content of individuals' hate speech definitions, confounding variables need to be accounted for. Dealing with educational attainment and political knowledge or opinions, the literature suggests taking into account *gender* (Costello et al., 2019; Cowan & Khatchadourian, 2003; Wilhelm & Joeckel, 2019; Wojatzki et al., 2018), *age* (Lambe, 2004), *political interest* (Hall, 2018), *political ideology* (Munzert et al., 2023), *social media usage* (Celuch et al., 2022; Costello et al., 2019), and *experience with hate speech* (Costello et al., 2019; Soral et al., 2018). Apart from social media usage, all these variables are included in the data set. Social media usage was not measured and hate speech experience is supplemented by experience with *online hostile engagement*.

The Directed Acyclic Graph (Fig. 1) results from the temporal sequence of social influences and effects already established in the literature and serves as the basis for the variable setup in the analysis. The sociodemographic variables *gender* and *age* are clearly biographically prior to the influences of formal education and therefore serve as control variables. A measure of *survey commitment* is used to reduce bias due to non-engagement. The treatment of political interest and ideology, as well as experience with hate speech and online hostile engagement is somewhat more complex, as it is certainly possible that formal education influences these characteristics. Therefore, *academic status* is used not only as a primary independent variable, but also as a moderator to examine its potential influence on the relationships between these variables and respondents' hate speech definitions.

The moderation analysis may provide further evidence of the role of academic education on the form and content of hate speech definitions. In particular, the

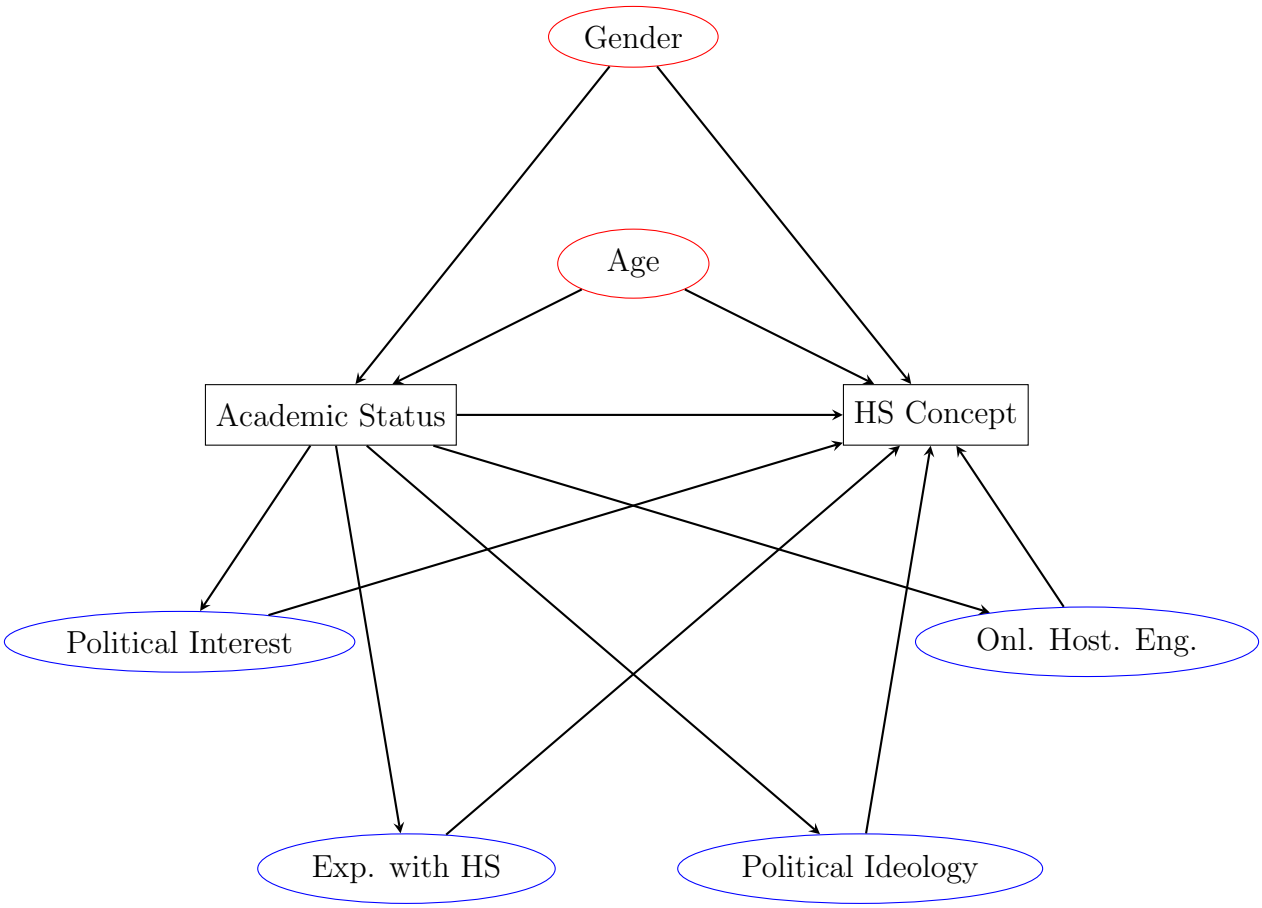


Figure 1: Directed Acyclic Graph with academic status as treatment variable and hate speech concept (including text length, readability, and content type) as outcome. Confounding variables are displayed in red and mediators in blue.

interaction of academic status with political ideology may provide insight into how hate speech definitions are politically shaped in relation to academic status.

3 Data

The data used in this analysis were collected for the study "Global Preferences for Hate Speech Moderation" (Munzert et al., forthcoming), an intervention experiment conducted as a cross-sectional survey in eleven countries (Brazil, Colombia, Germany, India, Indonesia, Nigeria, Philippines, Poland, Turkey, United Kingdom, United States) across selected regions with different cultural contexts and varying free speech norms. Participants were recruited through Facebook ads and targeted to be balanced across age (18-40 and 41-65+), gender (male, female), and education (below college, at least college). The original dataset includes responses from 19,172 respondents.

Next to a pretreatment survey, the dataset contains the results of a vignette experiment on hate speech regulation preferences, a normative framing experiment, and a question-ordering experiment. This study uses only pre-treatment variables extracted from survey items asked before the experiments, and an open-text item about people's individual definition of hate speech asked after the experiments. In addition, processing time metadata is used to measure respondents' commitment in answering the questions.

3.1 Hate Speech Definitions

The following open text item from the survey serves as the basis for creating representative indicators of respondents' individual definitions of hate speech:

"People have different ideas about what constitutes 'hate speech.' What about you - how would you personally define hate speech?"

In conducting the original study, the open-text responses to the hate speech definition question were partially classified according to common themes in hate speech definitions from the literature and patterns within the data. The coding scheme (see Appendix 8) for this task contained items on content, sender features and motivation, target scope features, specified target features, and other features of the statement (like, e.g., if the answer provides an example or no definition at all).

A manually annotated portion of the data ($n = 1721$) is provided as a basis for this work by the researchers conducting the main study. The dataset consists of 46 binary variables (features of definitions) indicating whether a specific feature (e.g. attack as content of hate speech) is existent in a definition ("1") or not ("0"). This annotated dataset also defines the selection and size of the analytical sample used for the analysis. For the purpose of meaningful clustering and the prediction of political orientation, observations with annotations flagged for revision or without any chosen item were filtered out, reducing the sample to $n = 1,549$. After randomly selecting one annotation per observation, the sample consists of 1,079 observations from Brazil (6%), Colombia (2%), United Kingdom (15%), Germany (16%), Indonesia (6%), India (9%), Nigeria (9%), Phillipines (9%), Poland (4%), and the United States (23%). After listwise deletion of missing observations in key variables, the final analytical sample consisted of 1,049 observations.

3.2 Formal Education

The respondents' level of formal education serves first as an independent variable and then as an interaction term in the regression analysis. The survey item measuring educational attainment is measured on a scale of 6 to 9 levels per country, depending on the specific structure of the different education systems. Based on a coarser categorization across countries, an additional three-level education variable was already created within the main study:

- **Low**: did not finish school (yet), or finished school but holds no qualification to pursue education to satisfy university entrance requirements
- **Intermediate**: finished school with qualification to pursue further education to satisfy university entrance requirements
- **High**: finished school achieving university entrance requirements, and/or holds university degree and/or post-graduate degree

To analyze differences between respondents with academic and non-academic backgrounds, a third educational variable *academic status* is constructed in this analysis, which summarizes low and medium levels as "non-academic" and high levels as "academic".

Educational levels vary by country and only partially overlap (Schneider, 2022). Schneider notes that harmonizing these levels post-survey is challenging, potentially distorting analysis results and hindering hypothesis testing on the effects of specific types of education, like vocational courses.

The distinction made for this study creates a comparable classification and compensates for inaccuracies in the measurement between low and intermediate education. The additional variable aims to divide the sample into individuals with at least a bachelor's degree according to the International Standard Classification ISCED levels 6, 7 and 8 (ISCED, 2012), who are assigned the status "academic". All lower educational qualifications, including primary, lower secondary, upper secondary, and short tertiary, are classified as "non-academic". Unlike the ISCED (2012), this study also considers a certain level of postgraduate or vocational education as "academic". This approach takes into account the socialization effects of spending time at a higher education institution, independent of an obtained degree.

3.3 Measurement of Covariates

Most of the suggested control variables were measured through the pretreatment survey. *Gender* (male/female/other) and *age* (year of birth, derived age, scale 1-10, divided into categories 18-29, 30-49, 50-69, 70+) were taken in their format from the original study. *Political interest* was measured with four levels and collapsed into three (low/medium/high) to achieve a more balanced distribution. *Political ideology* was first measured on a scale from 1 to 11 (1 = left and 11 = right) and then broken down into three categories of political spectrum (1-4 → liberal/left, 5-7 → moderate/center, 8-11 → conservative/right). For *experience with hate speech* and *online hostile engagement*, a score was calculated from the five and three questionnaire items, respectively.

As an additional control variable, timestamp variables were used as a measure of *survey commitment*. All time intervals spent on the previous components of the survey before the hate speech definition question were summed up. The pre-survey commitment was then assigned as numeric variable in minutes rounded to one decimal place, as well as minutes log transformed with base 2 due to the large accumulation of low values, and as a categorical variable divided by the mean of the logged minutes ("Above average"/"below average") for the sample descriptives and missingness analysis.

To measure how well the respondents' political attitudes can be predicted from their hate speech definitions, the following survey item was used: *"In politics people often talk about the "left" and the "right". On a scale between 1 (furthest left) and 11 (furthest right), where would you place yourself?"* with self-assessed values between 1 and 11. Descriptive statistics of all numeric and categorical variables are listed in the Appendix 4 and 5.

3.4 Missingness

Overall Missingness: 3728 out of 19,172 respondents did not answer the survey question about their personal definition of hate speech. This corresponds to 19.4 percent of the overall sample, which motivates a look at the characteristics of the respondents with missing answers and raises the question of whether there is non-response bias to consider when analyzing existing responses.

Missing respondents tended to show below-average survey commitment in general, but even of those with above-average survey commitment, 17 percent did not answer the question, indicating that it only explains a fraction of the reasons for high missingness (see Fig. 7 in the Appendix).

Respondents were also less likely to answer the open-ended question if they were in the 18-29 age group (25% missing in this subgroup), had no minority status (20%), low education (23%), low political interest (28%), and reported less experience with hate speech (21%) or hostile online engagement (24%). Missingness also varied greatly between countries, with the highest proportion in Turkey and India with more than 30 percent missing responses and the lowest in the United States, the United Kingdom and Brazil with less than 15%.

While respondents without hate speech definitions could not be included in the further analysis, for the other variables, missingness was handled partly by list-wise deletion (main variables) and partly by assigning NA's to the part of the data that was used as a training set for predicting political orientation. In order not to lose too much data, observations with missing values in the political prediction indicator were kept unless the specific analysis required them to be omitted. In the classification data provided, any rows and columns that are only zero were filtered out.

4 Methods

4.1 Measuring Form and Content of Open-text Answers

Four indicators were constructed to assess respondents' definitions of hate speech. *text length* and *readability* are measured to determine the formal characteristics of the response. Clustering is used to identify different *content types*. The fourth indicator measures the *predictability of political orientation* to examine how strongly the respondents' political attitudes are reflected in their text. The two form indicators *text length* and *readability* are based on the original text items. The two content-related indicators, *content type* and *predictability of political orientation*, are created using the annotations of the open-text responses. For distributions and correlations of the indicators see Figure 9 in the Appendix.

4.1.1 Text Length

The basic R function *nchar*() was used to measure the length of each observation. The calculation of the characters includes punctuation and spaces. Due to a high density in low values, the variable has been logarithmically transformed with a base of 2.

4.1.2 Readability

The R library *readability* (Rinker, 2022), a collection of readability tools that utilize the *syllable* package (Rinker, 2024), was used to measure the *readability* of respondents' hate speech definitions. It calculates the Flesch Kincaid Flesch (1948), Gunning Fog Index Gunning (1968), Coleman Liau Coleman and Liau (1975), SMOG Harry and Laughlin (1969), Automated Readability Index Smith and Senter (1967), and an average of the five readability scores. While the formulas are known to strongly correspond (van Oosten et al., 2010), the average score turned out to be suitable because it compensated for individual extreme values and inaccuracies in

the algorithms (see supplementary material *2-readability-descriptives* for the comparison of algorithms[1]).

This traditional approach has been used to pragmatically measure readability in different languages. Compared to more recent measures of readability in political language, these algorithms come from the perspective of educational research and applied psychology (Benoit et al., 2019), which is suitable for the case at hand.

4.1.3 Content Type

In the next step, different *content types* of hate speech definitions were defined using an unsupervised approach on all available manual annotated open-text responses. To do this, the *Partitioning Around Medoids (PAM)* clustering algorithm was applied to the binary classification data (derived from the coding scheme in Appendix 8) using the *cluster library* (Maechler et al., 2023). With this method described in Kaufman and Rousseeuw (2009), the identified medoids are guaranteed to be real data points, which makes the interpretation of clusters in the context of the binary attributes at hand more meaningful and intuitive. PAM clustering is also the method of choice because the binary variables have a skewed distribution of 0 and 1.

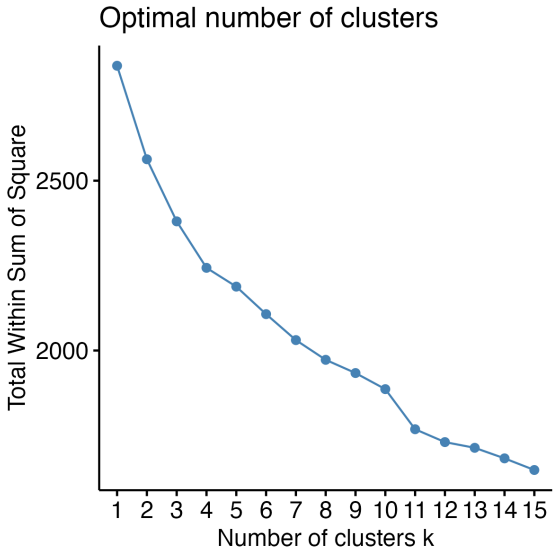


Figure 2: The total within-cluster sum of squares (WSS) by number of clusters

To determine an appropriate number of clusters, the elbow method was used. It plots the total within-cluster sum of squares (WSS) against the number of clusters, k. By looking for a "knee" in the plot where the rate of decrease changes sharply, the best number of clusters can be identified. The plot shows a large decrease up to a number of 4 clusters. A higher number of clusters did not improve the WSS much more (Fig. 2).

As there is no generally accepted technical or legal definition of the phenomenon, the respondents' definitions could not be compared with an "official" version. However, the clusters were explored and categorized based on the literature in 2.2. Taking into account the most frequent components, the clusters could be assigned to one of the three definition types according to Marwick and Miller (2014). As a result, one cluster can be described as intent-based and two others as harms-based. The fourth cluster represents respondents who either wrote nonsense or tend to deny the existence of an objective definition. One of the clusters in which hate speech is rather defined as harms-based can also be described as "harms-based narrow" due to a higher average number of classification elements based on the framework in Kansok-Dusche et al. (2023). More detailed results of the cluster analysis are listed in Table 6 in the Appendix.

4.1.4 Predictability of Political Orientation

How well does a respondent's personal definition of hate speech predict their political orientation? The methodological basis is a study by Hemphill et al. (2016), which measured how easy it is to predict a politician's political affiliation based on the individual's use of certain hashtags on X (formerly Twitter). The annotated hate speech definition data with the binary item variables (indicating the presence of certain content characteristics, target characteristics, etc. in the definitions) in combination with the outcome variable *leftright* was used to develop a linear model to predict the political spectrum.

It was decided to train the model on 20% of the dataset, as this yielded better results than a smaller fraction, but also results comparable to those obtained with larger training sets, thus providing an optimal balance between preserving the size of the dataset for analysis and ensuring robust model performance.

The model was then used to predict the left-right orientation of the remaining 80 percent of the data. The new variable *leftright prediction error* was created from the respective errors of this prediction and serves as another content indicator of the hate speech definitions. It quantifies the absolute difference between the predicted and actual political orientation, with lower values indicating a better predictability and thus a greater representation of political attitudes within the person's hate speech definition.

4.2 Statistical Analysis

To test for potential differences in sophistication when defining hate speech between academic status groups (H1), the two form indicators *text length* and *readability* of hate speech definitions are evaluated, as well as the distribution of *content types*.

Firstly, an independent samples t-test is used to compare the means of the numeric indicators. In order to gain a more nuanced understanding of the role of academic status, two linear regression models with *text length* and *readability* as outcomes were employed. The control variables *gender*, *age*, and *survey commitment* are included in the model to adjust for potential confounders (see DAG in 2.4).

To identify differences in the content of the definitions between both groups, a chi-square test for independence is used to test for a potential relationship between academic status and cluster affiliation. In addition, the definitions of academics and non-academics are compared by the most common characteristics and features of the content types that are potentially more typical of one group than another. A multinomial logistic regression serves to model the relationship between academic

status as independent variable and the content type represented through the four clusters as outcome variable, using the same controls as for the numeric indicators.

Finally, the effect of academic status as a moderator on the relationship between other important variables related to the definition of hate speech is analyzed. For this purpose, *academic status* is used as an interaction term with *political interest*, *experience with hate speech*, and *online hostile engagement* in the same three regression setups and plotted using the *sjPlot library* described in Lüdecke (2024).

In order to examine the political ideology in the definitions (H2), a first step is to compare the mean prediction error derived from the predictions of political orientation between the groups using an independent samples t-test. In addition, the other indicators are modeled using academic status as an interaction term with political ideology to provide further evidence of the role of political ideology in how different academic status groups define hate speech.

5 Results

5.1 Comparing Indicators between Academic Status Groups

5.1.1 Significance Tests

To analyze differences in personal hate speech definitions between academics and non-academics, several indicators of form and content were compared between the groups. Text length, measured in number of characters, showed no significant differences ($t = -0.27$, $df = 967.04$, $p = 0.79$). This suggests that the amount of text written by academics ($mean = 131$) is similar to that of non-academics ($mean = 133$). A significant difference was found for Readability ($t = 2.19$, $df = 990.53$, $p = 0.029$), with academics' texts being slightly more readable with a mean score of 11.88 compared to 11.15 for non-academics ($SD = 5.5$). The predictability of political orientation based on the prediction error showed no significant differences between the

groups ($t = -0.125$, $df = 781.05$, $p = 0.900$), indicating that in this sample, political orientation is similarly predictable from the hate speech definitions of both academics and non-academics.

5.1.2 Differences in Content types

The frequencies in content types of hate speech definitions varied significantly between academics and non-academics, as indicated by the chi-squared test ($\chi^2 = 13.391$, $df = 3$, $p = 0.004$). The distribution of the academic status groups across the four identified content types is shown in Figure 3. A general trend can be seen that both groups tend to define hate speech as either harms-based or intent-based, whereby academics are roughly equally distributed in these two clusters, but non-academics define the phenomenon harms-based much more frequently than intent-based. In addition, academics define hate speech more frequently in harms-based and narrow terms (naming many attributes) than non-academics. They, on the other hand, were more frequently found in the group of those who denied the existence of the phenomenon or answered nonsense.

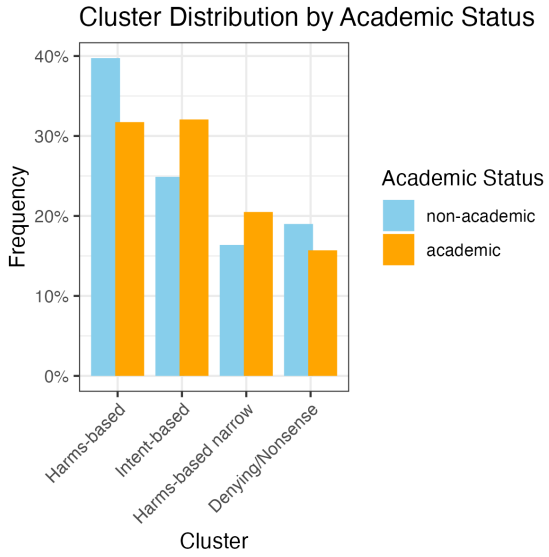


Figure 3: Distribution of academic status groups across clusters (percentage of the respective academic status group)

To gain a better understanding of the significant differences in content types, typ-

ical components of the definitions can be examined. Table 1 shows that the five most frequent components in the hate speech definitions in both groups include "Scope People", "Scope Person", "Content Attack" and "Content Incitement", which occur with only slight differences in frequency. Apart from that, one item in the list differs: "Target Race" occurs in 16 percent of the definitions of academics and "Nonsense" in 19 percent of the definitions of non-academics.

Academic		**Non-academic**	
Component	**Percentage**	**Component**	**Percentage**
Scope People	35%	Content Attack	32%
Scope Person	33%	Scope People	28%
Content Attack	30%	Scope Person	25%
Content Incitement	21%	Nonsense	19%
Target Race	16%	Content Incitement	16%

Table 1: The five most frequently identified items per academic status group and the percentage of their occurrence in this group.

5.1.3 Predicting Political Orientation

The performance metrics of the prediction ($RMSE = 2.756$, $R^2 = 0.047$, $MAE = 2.139$) reflect moderate predictive ability, but with a low power to explain the variance in political orientation. Importantly for the further handling of the indicator, no significant correlation ($\rho = .041$) was found between the original leftright distribution and the prediction errors, suggesting unbiased prediction across the political spectrum.

Evaluating the feature importance, most important predictors pointing towards a certain political spectrum were about the *target* of hate speech: age ($\beta = 4.458$), class ($\beta = 3.739$), region ($\beta = 2.427$) leading to a prediction more to the right, while sexual orientation ($\beta = -1.785$), tradition ($\beta = -1.599$), and vulnerable groups ($\beta = -1.574$) leading to a prediction more towards the left side of the political spectrum.

For a comprehensive insight into the predictive features and model accuracy, de-

scriptives of the highest and lowest scores, as well as the full list of feature importance within the model, see additional material *4-leftright-prediction-descriptives.html*[1]. The features used for model training are taken from the annotation dataset based on the coding scheme for hate speech definitions (Appendix 8).

5.2 The Effect of Academic Status on Hate Speech Definitions

5.2.1 Direct Effects

The first linear regression model (Tab. 2), predicting text length as logged to base 2, demonstrates that academic status shows no significant influence on *text length*. The coefficient is $\beta = 0.024$ with a standard error of $SE = 0.081$, indicating a minimal and statistically insignificant effect. However, academic status positively affects readability with a significant coefficient of $\beta = 0.758$ ($p < 0.05$), and a standard error of $SE = 0.334$, albeit only by a small fraction of one standard deviation (readability values ranging from -7.9 to 42.7 with $SD = 5.5$, see Tab. 4 in the Appendix). Conservative/right views are associated with shorter and less readable texts, while survey commitment is positively related with both outcomes.

The analysis of the influence of academic status on different content types of hate speech definitions (Tab. 3) revealed nuanced outcomes. Academic status significantly increases the likelihood of categorizing hate speech definitions as intent-based, with a positive coefficient ($\beta = 0.428$, reference catagory: Denying/Nonsense). Conversely, academic status did not significantly affect the propensity to categorize definitions as harms-based. For the harms-based narrow category, academic status again is a significant predictor with a positive coefficient ($\beta = 0.379$). Among the covariates, political ideology is once more found to play a significant and dominant role, this time together with age, which shows a notable influence on the distribution of clusters.

Table 2: Linear regression results showing marginal effects of academic status on text length and readability of hate speech definitions, controlled for gender, age, political ideology, and survey commitment. Standard errors are provided in parentheses, and significance levels are indicated.

	Text length (log2)	**Readability**
Academic status (ref = Non-academic)		
Academic	0.024	0.758**
	(0.081)	(0.334)
Gender (ref = Male)		
Female	−0.038	0.279
	(0.086)	(0.356)
Other	0.524	1.705
	(0.327)	(1.352)
Age (ref = 18-29)		
30-49	−0.171	0.206
	(0.111)	(0.459)
50-69	−0.204*	−0.755
	(0.113)	(0.467)
70+	−0.238	−2.308***
	(0.191)	(0.789)
Political ideology (ref = Liberal/left)		
Moderate/Center	−0.132	−1.763***
	(0.100)	(0.415)
Conservative/Right	−0.334***	−2.004***
	(0.110)	(0.454)
Survey commitment (log2)	0.297***	0.781***
	(0.067)	(0.277)
Constant	5.842***	10.130***
	(0.253)	(1.047)
Observations	1,049	1,049
R^2	0.034	0.049
Adjusted R^2	0.025	0.040
Residual Std. Error (df = 1039)	1.285	5.319
F Statistic (df = 9; 1039)	4.002***	5.909***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3: Multinomial regression results showing the impact of academic status on respondents' classification of hate speech definitions into the different content types, using denying/nonsense as the reference category, controlled for gender, age, political ideology, and survey commitment. Standard errors are provided in parentheses, and significance levels are indicated.

	Intent-based	**Harms-based**	**Harms-based narrow**
Academic status (ref = Non-academic)			
Academic	0.428**	−0.093	0.379*
	(0.198)	(0.187)	(0.216)
Gender (ref = Male)			
Female	0.472**	0.577***	0.642***
	(0.228)	(0.218)	(0.242)
Other	−0.385	−1.709**	−1.109
	(0.637)	(0.858)	(0.871)
Age (ref = 18-29)			
30-49	−0.236	0.026	−0.220
	(0.279)	(0.277)	(0.304)
50-69	−0.821***	−0.269	−0.568*
	(0.283)	(0.276)	(0.304)
70+	−0.528	0.154	−0.745
	(0.474)	(0.443)	(0.544)
Political ideology (ref = Liberal/left)			
Moderate/Center	−1.247***	−0.878***	−1.259***
	(0.304)	(0.302)	(0.322)
Conservative/Right	−1.866***	−1.367***	−1.589***
	(0.315)	(0.308)	(0.332)
Survey commitment (log2)	0.280*	0.089	0.269
	(0.168)	(0.162)	(0.182)
Constant	0.862	1.342**	0.299
	(0.654)	(0.635)	(0.705)
Akaike Inf. Crit.	2,778.805	2,778.805	2,778.805
Note:			*p<0.1; **p<0.05; ***p<0.01

5.2.2 Academic Status as Moderator

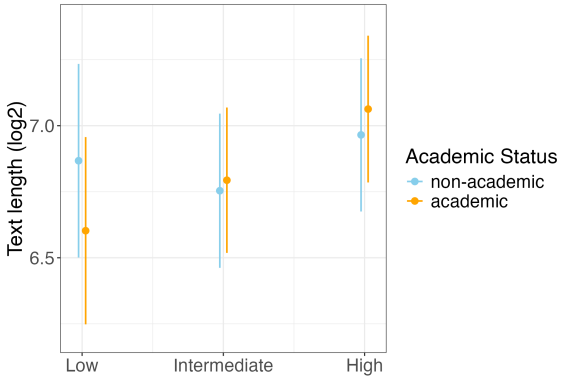
While academic status showed no significant effect on text length, it does play a role as moderator: While the *text length* of non-academics with different levels of political interest does not differ considerably, Figure 4a shows a positive relationship among academics in a range of half a standard deviation ($SD = 1.2$). In the politically left-liberal camp, the predicted text length of academics is larger than those moderate/center, and the difference is even stronger towards conservatives/rights, whereas again there is no clear pattern for non-academics (Fig. 4b). Experience with hate speech increases the predicted text length similarly regardless of the academic status (Fig. 4c), but reported online hostile engagement leads to a higher predicted text length for academics, while it showed no effect for non-academics (Fig. 4d).

the impact of academic status on respondents' classification of hate speech definitions into the different content types, using denying/nonsense as the reference category, controlled for gender, age, political ideology, and survey commitment. Standard errors are provided in parentheses, and significance levels are indicated.

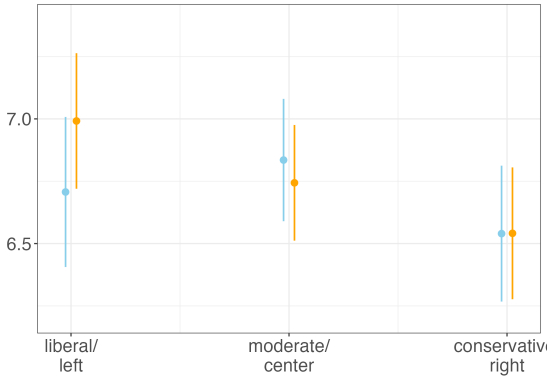
Readability shows a similar trend, as shown in Figure 5, with the highest predicted differences being around half a standard deviation ($SD = 5.5$) for academics in political ideology (Fig. 5b). In this case, for non-academics, more online hostile engagement is associated with a lower predicted readability score, while the pattern for academics did not change (Fig. 5d).

Figure 6 visualizes the interaction effects between academic status and important covariates on the content type of definitions. In the case of political interest, the patterns appear to be similar by academic status. However, non-academics with intermediate political interest are more likely to define hate speech harms-based than academics, while the latter are more likely to define the phenomenon as harms-based narrow than those without academic education (Fig. 6a).

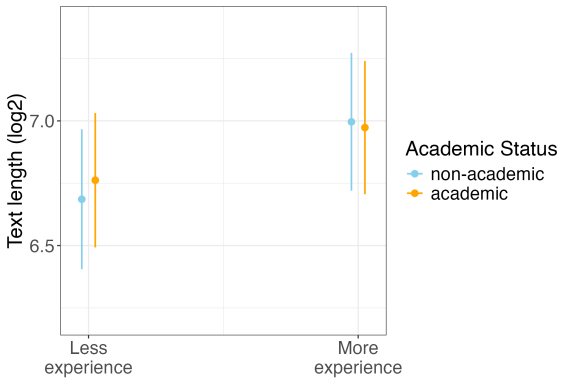
In the case of political ideology (Fig. 6b), academics in the moderate/center spectrum are more likely to be found in the intent-based cluster than non-academics



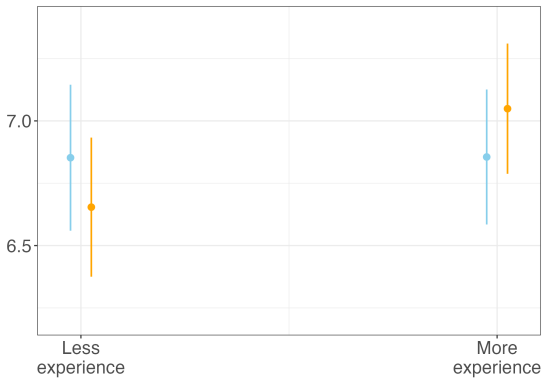
(a) Political Interest



(b) Political Ideology

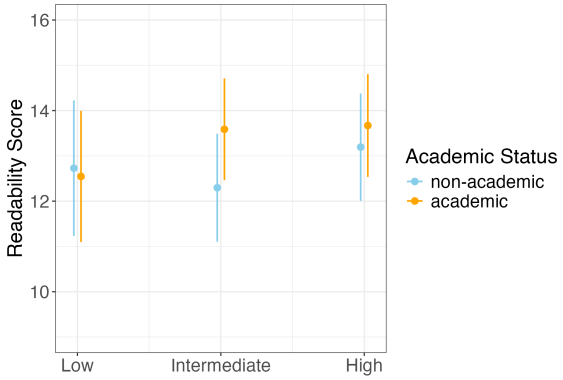


(c) Experience with HS

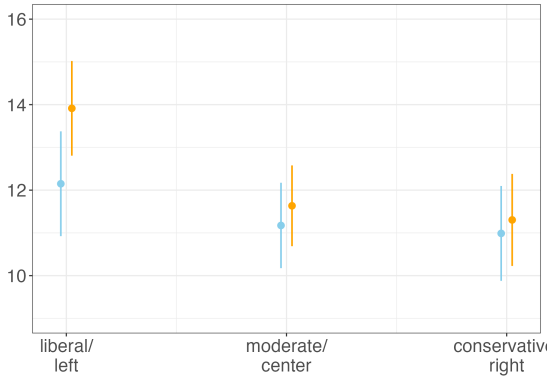


(d) Online Hostile Engagement

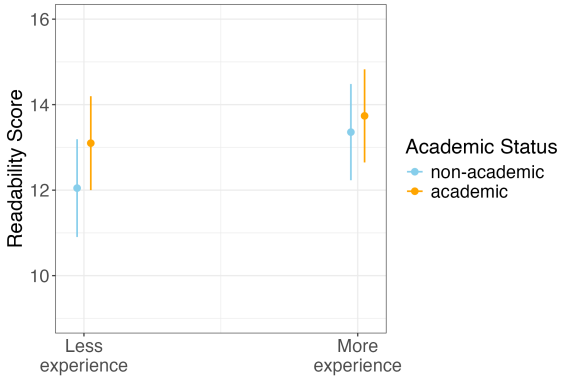
Figure 4: Predicted values (marginal effects) for text length by academic status interacting with political interest, political ideology, experience with hate speech, and online hostile engagement. Controlled for gender, age, and survey commitment.



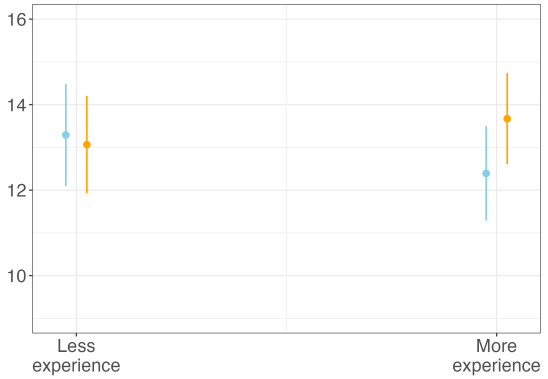
(a) Political Interest



(b) Political Iedology



(c) Experience with HS

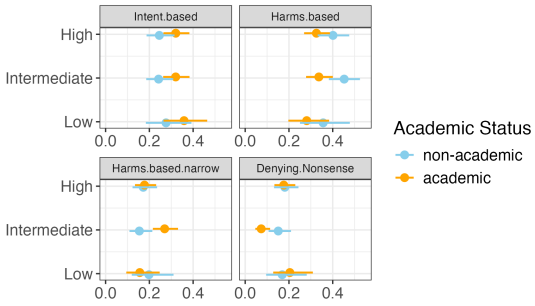


(d) Online Hostile Engagement

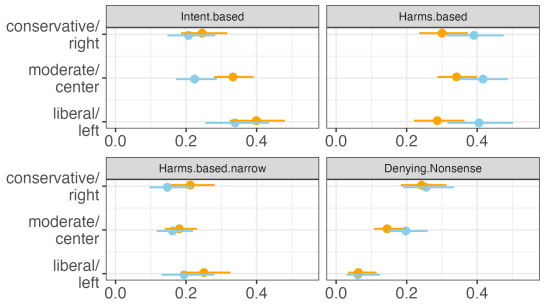
Figure 5: Predicted values (marginal effects) for readability by academic status interacting with political interest, political ideology, experience with hate speech, and online hostile engagement. Controlled for gender, age, and survey commitment.

with the same political ideology. The level of experience with hate speech among non-academics indicates slightly less impact on the content type of their definition than the experience of academics (Fig. 6c).

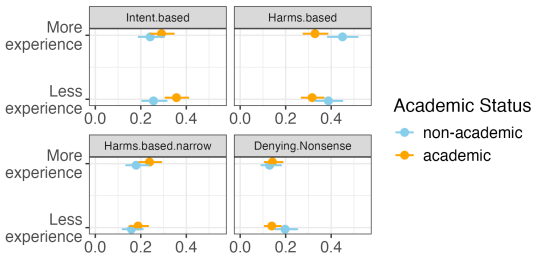
The largest group differences can be observed in the interaction with online hostile engagement (Fig. 6d). More experience among academics leads more likely to an intent-based definition than less experience, while it stays constant for non-academics. The predicted probability of a harms-based definition increases with more experience, while it decreases for academics with more experience. For the two clusters *harms-based narrow* and *denying/nonsense*, there is no notable difference in the effect between the groups.



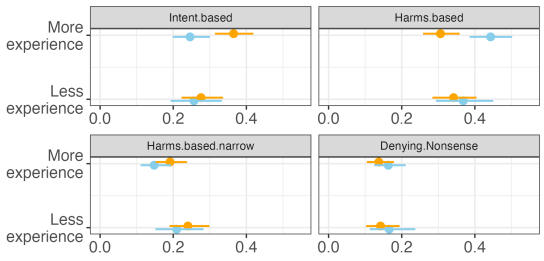
(a) Political Interest



(b) Political Ideology



(c) Experience with HS



(d) Online Hostile Engagement

Figure 6: Estimated marginal effects for content types by academic status. Controlled for gender, age, and survey commitment. Other predictors are held constant (discrete predictors are held constant at their proportions, not the reference level)

6 Discussion

The present study investigated whether the academic status of respondents influences the form and content of their hate speech definitions. It can not be confirmed that the definitions of academics are generally longer or remarkably more readable, which speaks against H1. However, academics tend to define hate speech more intent-based or narrowly harms-based than non-academics. This indicates a more abstract and detailed understanding of the concept and thus supports the assumption of more sophisticated definitions of academic respondents (H1). Narrower definitions could also point to better knowledge about specific target groups.

The definitions of non-academics in the sample are more often broad, considering hate speech as subjective or negating the existence of the phenomenon. This approach does not necessarily express a lack of knowledge: Similar to the finding of Jubany and Roiha (2016), this suggests that most people possess an intuitive understanding of the phenomenon. However, as described in Munzert et al. (2023), political ideology determines the assessment of what is considered hate speech and what is not. Age was also a factor found more important than academic status, with a higher age indicating shorter, less readable texts and for those between 50 and 69 years, a higher probability to deny the existence of hate speech. Thus, it can be concluded that academic status is not a strong factor in the formation of longer or more readable definitions within the analyzed sample. Instead, if there are tendencies, they are mostly outweighed by the influence of political attitudes, age, or gender.

Similarly, a comparison of political prediction errors by academic status group alone demonstrated no relevant differences, suggesting that it is not easier to predict the political spectrum from definitions of academics than non-academics. Thus, definitions of the former seem not more politically defined (H2). Since the model makes fairly reliable predictions across groups, it is, however, worth noting that in both directions of the political spectrum, specific target features were more impor-

tant predictors than other components of the definitions, clearly indicating political views on who is most affected by hate speech: *age* or *class* as targets lead to a more conservative/right-wing prediction, and *sexual orientation* or *vulnerable groups* to left/liberal, supporting the finding of a strong influence of political views on definitions compared to formal knowledge.

The moderator analysis revealed a nuanced picture of how political interest, political attitudes and experiences with hate speech shape the content of the definitions differently depending on academic status. As hate speech is a "fundamental normative issue" (Munzert et al., 2023), it is not surprising that definitions of the concept are based on political views. However, the definitions of academics with a higher political interest and orientation on the left of the political spectrum have a significantly greater text length and readability compared to those with lower interest and further to the right, while there is no clear pattern for non-academics in either context. This suggests that the knowledge of hate speech or the effort academics invest in their response is more strongly related to their political interest and political attitudes than for non-academics. This supports the assumption that political attitudes play a greater role in defining hate speech for academics than for non-academics (H2).

The findings of the study must be considered with several limitations in mind. Primarily, the study's sample was not representative, relying on self-selection and limited to Facebook users within specific groups. This limitation emphasizes the absence of inferential knowledge, as generalizations to the broader population are constrained. Furthermore, academic status groups were not weighted or matched in their composition, which could be a promising approach for future research.

The missingness analysis (Fig. 7 in the Appendix) revealed a clear pattern indicating that individuals with lower levels of education and political interest, as well as those with less experience with hate speech or online hostile engagement, were less likely to respond to the question about the definition, resulting in their exclusion

from the sample. Notably, approximately 25 percent of the 18-29 age group also did not respond, further distorting the sample. The sample is particularly imbalanced with regard to gender and age (see Tab. 5 in the Appendix).

It is also important to note that participants were exposed to hate speech examples prior to the question item on defining hate speech. This could have primed them with a certain awareness of what answers are socially desirable and simply provided them with more knowledge about the phenomenon. It could weaken the educational effect somewhat and make ideological attitudes more prominent, especially if certain political camps feel incited by the examples they see to express their political opinions. This is particularly notable among academics, who may possess a greater understanding of societal expectations in online behavior and reconcile new examples of hate speech more quickly with their knowledge base and political views.

Furthermore, unobserved factors such as the content of the participants' training could influence the results more significantly than their educational degrees — for instance, whether someone studied engineering or political sciences. The measurement of variables like empathy and online hostile engagement may be biased due to self-assessment.

The study did not account for country-specific differences, which could impact the interpretation and evaluation of hate speech. For future research, it would be beneficial to explore the effects of country-specific educational systems on hate speech perceptions, particularly those with an emphasis on democratic citizenship education. This could provide insights into differences between societies in defining hate speech.

Studying these aspects can provide a more nuanced understanding of the dynamics of people's hate speech definitions and help in developing better regulatory frameworks. As Sellars (2016, p. 32) notes, *"The definition of hate speech in a study or regulatory environment may be the most important part of the project's design."*

While the impact of education per se remains uncertain, the political struggle to

foster a culture of accountability and mutual respect in digital spaces certainly remains reliant upon educational efforts. Promising approaches include public awareness campaigns, targeted training to improve the detection and reporting of hate incidents online, and the promotion of internet awareness and counter-speech among students (Jubany & Roiha, 2016). In particular, democratic citizenship education addresses not only knowledge about hate speech, but also its normative aspects (Estellés & Castellví, 2020; Keen et al., 2020; United Nations, 2019), which have been shown to be influential in shaping people's individual definitions of hate speech.

References

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and Explaining Political Sophistication through Textual Complexity. *American Journal of Political Science*, *63*(2), 491–508. https://doi.org/10.1111/ajps.12423

Bobo, L., & Licari, F. C. (1989). Education and Political Tolerance: Testing the Effects of Cognitive Sophistication and Target Group Affect. *Public Opinion Quarterly*, *53*(3), 285–308. https://doi.org/10.1086/269154

Celuch, M., Oksanen, A., Räsänen, P., Costello, M., Blaya, C., Zych, I., Llorent, V. J., Reichelmann, A., & Hawdon, J. (2022). Factors Associated with Online Hate Acceptance: A Cross-National Six-Country Study among Young Adults. *International Journal of Environmental Research and Public Health*, *19*(1), 534. https://doi.org/10.3390/ijerph19010534

Ceobanu, A. M., & Escandell, X. (2010). Comparative Analyses of Public Attitudes Toward Immigrants and Immigration Using Multinational Survey Data: A Review of Theories and Research. *Annual Review of Sociology*, *36*(1), 309–328. https://doi.org/10.1146/annurev.soc.012809.102651

Chan, J. (2019). The Effect of College Education on Intolerance: Evidence from Google Search Data. *Applied Economics Letters*, *26*(2), 83–86. https://doi.org/10.1080/13504851.2018.1438582

Chan, J. M., Chau, K. K. L., & Lee, F. L. F. (2002). Abstract Principle Versus Concrete Interest: A Study of Education and Political Opinion in Hong Kong. *International Journal of Public Opinion Research*, *14*(1), 54–72. https://doi.org/10.1093/ijpor/14.1.54

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284. https://doi.org/10.1037/h0076540

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social Group Identity and Perceptions of Online Hate*. *Sociological Inquiry*, *89*(3), 427–452. https://doi.org/10.1111/soin.12274

Council on Foreign Relations. (2019, July). *Hate Speech on Social Media: Global Comparisons* (tech. rep.). Retrieved May 5, 2024, from https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons

Cowan, G., & Khatchadourian, D. (2003). Empathy, Ways of Knowing, and Interdependence as Mediators of Gender Differences in Attitudes Toward Hate Speech and Freedom of Speech [Publisher: SAGE Publications Inc]. *Psychology of Women Quarterly*, *27*(4), 300–308. https://doi.org/10.1111/1471-6402.00110

Estellés, M., & Castellví, J. (2020). The Educational Implications of Populism, Emotions and Digital Hate Speech: A Dialogue with Scholars from Canada, Chile, Spain, the UK, and the US. *Sustainability*, *12*(15), 6034. https://doi.org/10.3390/su12156034

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, (Bd. 32, Ausg. 3).

Guillén-Nieto, V. (2023). *Hate Speech: Linguistic Perspectives*. De Gruyter. https://doi.org/10.1515/9783110672619

Gunning, R. (1968). *The technique of clear writing*. New York, McGraw-Hill. Retrieved May 16, 2024, from http://archive.org/details/techniqueofclear00gunn

Hall, J. P. (2018). Effects of Political Knowledge on Political Tolerance [Publisher: Routledge _eprint: https://doi.org/10.1080/15512169.2017.1366326]. *Journal of Political Science Education*, *14*(1), 104–122. https://doi.org/10.1080/15512169.2017.1366326

Harry, G., & Laughlin, M. (1969). SMOG Grading - A New Readability Formula. *The Journal of Reading*.

Heijden, E. v. d., & Verkuyten, M. (2020). Educational Attainment, Political Sophistication and Anti-Immigrant Attitudes. *Journal of Social and Political Psychology*, *8*(2), 600–616. https://doi.org/10.5964/jspp.v8i2.1334

Hemphill, L., Culotta, A., & Heston, M. (2016). #Polar Scores: Measuring partisanship using social media content. *Journal of Information Technology & Politics*, *13*(4), 365–377. https://doi.org/10.1080/19331681.2016.1214093

Highton, B. (2009). Revisiting the Relationship between Educational Attainment and Political Sophistication [Publisher: The University of Chicago Press]. *The Journal of Politics*, *71*(4), 1564–1576. https://doi.org/10.1017/S0022381609990077

ISCED, U. (2012). International standard classification of education 2011. *UNESCO Institute for Statistics*.

Izquierdo Montero, A., Laforgue-Bullido, N., & Abril-Hervás, D. (2022). Hate speech: A systematic review of scientific production and educational considerations. *Revista Fuentes*, *2*(24), 222–233. https://doi.org/10.12795/revistafuentes.2022.20240

Jubany, O., & Roiha, M. (2016). *Backgrounds, experiences and responses to online hate speech: A comparative cross-country analysis* (tech. rep.). Universitat de Barcelona. http://www.unicri.nu/special_topics/hate_crimes/Backgrounds_Experiences_and_Responses_to_Online_Hate_Speech_A_Comparative_Cross-Country_Analysis.pdf

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A Systematic Review on Hate Speech among Children and Adolescents: Definitions, Prevalence, and Overlap with Related Phenomena [Publisher: SAGE Publications]. *Trauma, Violence, & Abuse*, *24*(4), 2598–2615. https://doi.org/10.1177/15248380221108070

Kaufman, L., & Rousseeuw, P. J. (2009, September). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Keen, E., Georgescu, M., & Gomes, R. (2020, May). *Bookmarks (2020 Revised ed): A manual for combating hate speech online through human rights education.* Council of Europe.

Lambe, J. L. (2004). Who Wants to Censor Pornography and Hate Speech? *Mass Communication and Society*, *7*(3), 279–299. https://doi.org/10.1207/s15327825mcs0703_2

Lüdecke, D. (2024). Plotting Interaction Effects of Regression Models. Retrieved May 15, 2024, from https://strengejacke.github.io/sjPlot/articles/plot_interactions.html

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K., & Murphy, K. (2023, December). Cluster: "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. Retrieved April 26, 2024, from https://cran.r-project.org/web/packages/cluster/index.html

Marwick, A. E., & Miller, R. (2014, June). Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape. Retrieved April 22, 2024, from https://papers.ssrn.com/abstract=2447904

Munzert, S., Traunmüller, Richard, Barberá, Pablo, Guess, Andrew, & Yang, JungHwan. (2023). Citizen Preferences for Online Hate Speech Regulation (Working Paper).

Parekh, B. (2012). Is There a Case for Banning Hate Speech? In M. Herz & P. Molnar (Eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (pp. 37–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139042871.006

Rinker, T. (2022, December). Readability: Fast readability scores for text data. Retrieved April 18, 2024, from https://github.com/trinker/readability

Rinker, T. (2024, February). Syllable: A Small Collection of Syllable Counting Functions [original-date: 2015-08-02T02:10:21Z]. Retrieved April 18, 2024, from https://github.com/trinker/syllable

Schneider, S. L. (2022). The classification of education in surveys: A generalized framework for ex-post harmonization. *Quality & Quantity*, *56*(3), 1829–1866. https://doi.org/10.1007/s11135-021-01101-1

Sellars, A. (2016, December). Defining Hate Speech. https://doi.org/10.2139/ssrn.2882244

Smith, E. A., & Senter, R. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*.

Solovev, K., & Pröllochs, N. (2022). Hate Speech in the Political Discourse on Social Media: Disparities Across Parties, Gender, and Ethnicity. *Proceedings of the ACM Web Conference 2022*, 3656–3661. https://doi.org/10.1145/3485447.3512261

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

United Nations. (1966, December). International Covenant on Civil and Political Rights. Retrieved April 22, 2024, from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

United Nations. (2012, May). The Rabat Plan of Action. Retrieved April 22, 2024, from https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action

United Nations. (2019). United Nations Strategy and Plan of Action on Hate Speech. Retrieved April 23, 2024, from https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml

van Oosten, P., Tanghe, D., & Hoste, V. (2010). Towards an improved methodology for automated readability prediction. *7th Conference on International Language Resources and Evaluation (LREC 2010)*, 775–782. Retrieved April 18, 2024, from http://hdl.handle.net/1854/LU-1055826

Weakliem, D. L. (2002). The Effects of Education on Political Opinions: An International Study. *International Journal of Public Opinion Research*, *14*(2), 141–157. https://doi.org/10.1093/ijpor/14.2.141

Weinschenk, A. C., & Dawes, C. T. (2019). The Effect of Education on Political Knowledge: Evidence From Monozygotic Twins [Publisher: SAGE Publications Inc]. *American Politics Research*, *47*(3), 530–548. https://doi.org/10.1177/1532673X18788048

Wilhelm, C., & Joeckel, S. (2019). Gendered Morality and Backlash Effects in Online Discussions: An Experimental Study on How Users Respond to Hate Speech Comments Against Women and Sexual Minorities. *Sex Roles*, *80*(7), 381–392. https://doi.org/10.1007/s11199-018-0941-5

Witschge, J., Rözer, J., & van de Werfhorst, H. G. (2019). Type of education and civic and political attitudes. *British Educational Research Journal*, *45*(2), 298–319. https://doi.org/10.1002/berj.3501

Wojatzki, M., Horsmann, T., Gold, D., & Zesch, T. (2018). Do Women Perceive Hate Differently: Examining the Relationship Between Hate Speech, Gender, and Agreement Judgments [Publisher: DuEPublico: Duisburg-Essen Publications online, University of Duisburg-Essen, Germany]. https://doi.org/10.17185/DUEPUBLICO/72288

Appendix

Table 4: Sample descriptives of numerical variables ($n = 1,049$ for all variables except left-right prediction error with $n = 840$)

Variable	**Mean**	**SD**	**Min**	**Max**
Age	45.1	15.6	18	88
Political interest	3.2	0.8	1	4
Left-right	6.2	2.5	1	11
Empathy	−0.03	1.3	−2.1	4.7
Experience with hate speech	3.0	0.5	1.4	4.0
Online hostile engagement	2.3	0.7	1.0	4.0
Survey commitment	11.3	6.6	2.6	100.9
Survey commitment (log 2)	3.4	0.6	1.4	6.7
Text length	137.5	140.4	1	1,271
Text length (log 2)	6.6	1.3	0.0	10.3
Readability	11.6	5.5	−7.9	42.7
Left-right prediction error	2.1	1.7	0.003	9.9

Table 5: Sample descriptives by academic status

	Non-academic ($n = 451$)	**Academic** ($n = 598$)	**Overall** ($n = 1,049$)
Gender			
Male	314 (69.6%)	370 (61.9%)	684 (65.2%)
Female	130 (28.8%)	219 (36.6%)	349 (33.3%)
Other	7 (1.6%)	9 (1.5%)	16 (1.5%)
Age			
18-29	92 (20.4%)	113 (18.9%)	205 (19.5%)
30-49	163 (36.1%)	241 (40.3%)	404 (38.5%)
50-69	173 (38.4%)	207 (36.4%)	380 (36.2%)
70+	23 (5.1%)	37 (6.2%)	60 (5.7%)
Political interest			
Low	71 (15.7%)	93 (15.6%)	164 (15.6%)
Intermediate	191 (42.4%)	240 (40.1%)	431 (41.1%)
High	189 (41.9%)	265 (44.3%)	454 (43.3%)
Empathy			
Less empathetic	220 (48.8%)	337 (56.4%)	557 (53.1%)
More empathetic	231 (51.2%)	261 (43.6%)	492 (46.9%)
Experience with Hate speech			
Less experience	243 (53.9%)	321 (53.6%)	564 (53.8%)
More experience	208 (46.1%)	277 (46.4%)	485 (46.2%)
Online hostile engagement			
Less experience	151 (33.5%)	255 (42.6%)	406 (38.7%)
More experience	300 (66.5%)	343 (57.4%)	643 (61.3%)
Survey commitment			
Below average	241 (53.4%)	348 (58.2%)	589 (56.1%)
Above average	210 (46.6%)	250 (41.8%)	460 (43.9%)

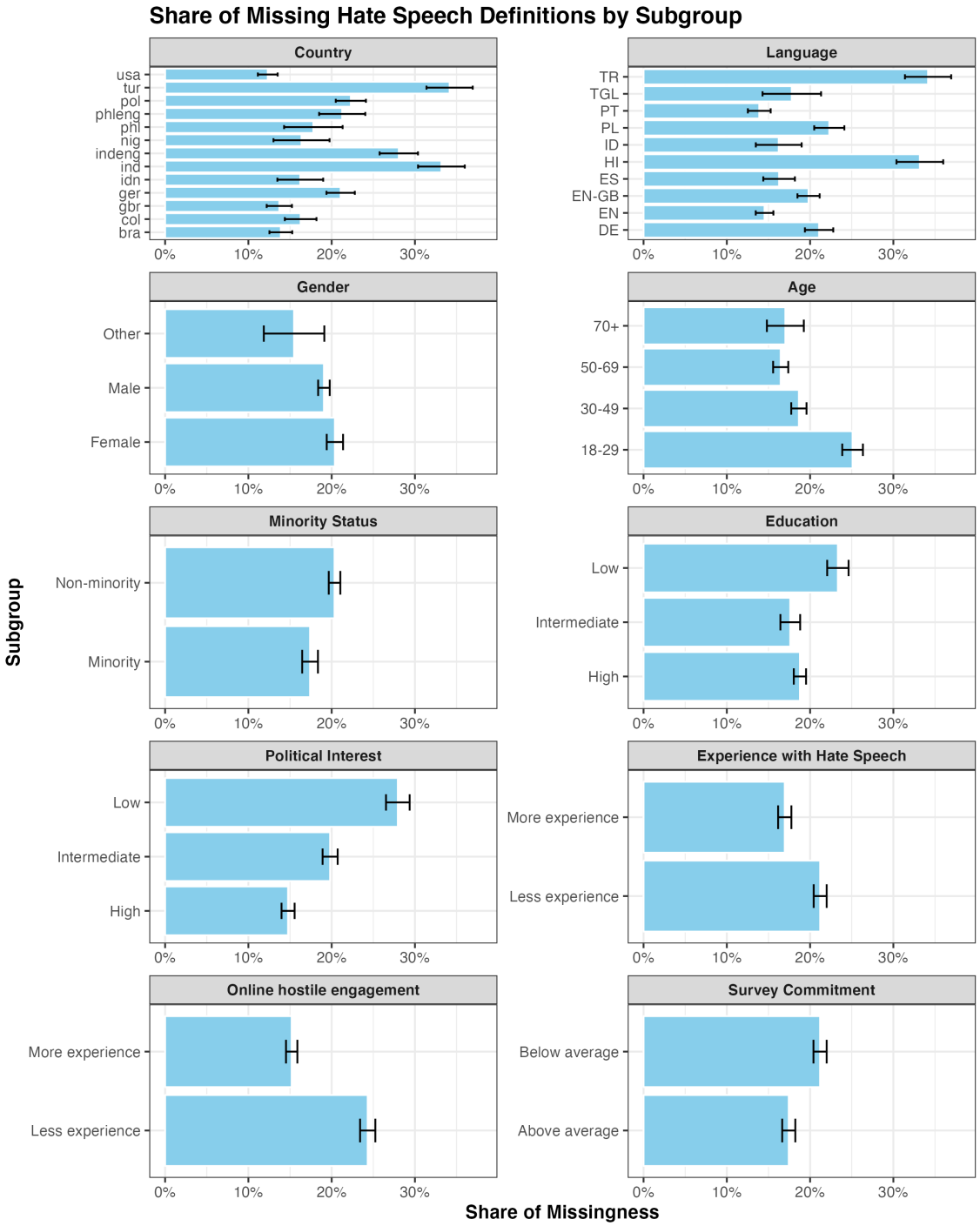


Figure 7: Share of missing answers by subgroup with a confidence level of 5% and sample size $n = 19,172$

Which of the following features of hate speech are mentioned in the definition?

Content
- ☐ offensive language
- ☐ hateful language
- ☐ severity
- ☐ mocking
- ☐ incivility
- ☐ harassment
- ☐ dehumanization
- ☐ attack
- ☐ opinion
- ☐ misinformation
- ☐ incitement
- ☐ causing harm
- ☐ against the law
- ☐ form of speech

Other content feature

Sender features/motivations
- ☐ who speaks generally matters
- ☐ person with power
- ☐ speaks in bad faith
- ☐ is prejudiced
- ☐ wants to hurt

Other sender feature

Target scope features
- ☐ against person/individual
- ☐ against group

Other scope feature

Specified target features
- ☐ specific features, general
- ☐ race/ethnicity/nationality
- ☐ region/geographical area
- ☐ sex/gender
- ☐ religion/belief system
- ☐ sexual orientation
- ☐ ideology/political affiliation
- ☐ class
- ☐ caste
- ☐ tribe
- ☐ age
- ☐ physical attributes
- ☐ profession/job
- ☐ culture/tradition
- ☐ minority, general
- ☐ vulnerable, general
- ☐ uncontrolled features, general
- ☐ more unspecified features (e.g., "etc.")

Other target feature

Other features of statement
- ☐ provides an example
- ☐ provides no definition
- ☐ provides negating definition
- ☐ questions/denies its existence
- ☐ questions its importance/relevance
- ☐ advocates for free speech
- ☐ emphasizes the subjectivity of hate speech
- ☐ elaborates on what should (not) be done against it

Other feature

Additional notes on classification

☐ Flag revision for further inspection

Elaborate on flag here (optional)

Figure 8: Coding scheme for hate speech definitions and basis for the annotation dataset

Table 6: Description of the clusters underlying the content types of hate speech definitions

Cluster Name	Frequency	Items per Definition (Variance)	Most frequent Items	Example Definition
Intent-based	29%	5 (4.3)	against group/people, incitement, against person/individual	*"Anything that uses falsehood, disinformation etc to target a person or group of people."*
Harms-based	35%	5 (3.0)	attack, hateful language, race as target	*"Negative and toxic content that doesn't help emotionally or productively"*
Harms-based narrow	19%	8 (3.3)	against person/individual, attack, against group/people, dehumanization	*"Direct threats, such as 'I will kill you and your family'. Just because someone doesn't like something does not make it hate speech."; "Threatening or hurting one's religious views and sentiments."*
Denying/Nonsense	17%	5 (0.4)	nonsense, denying existence/importance, hate speech is a subjective matter	*"There is no such thing"; "People should be free to say whatever they want even if if hurts another's feelings."*

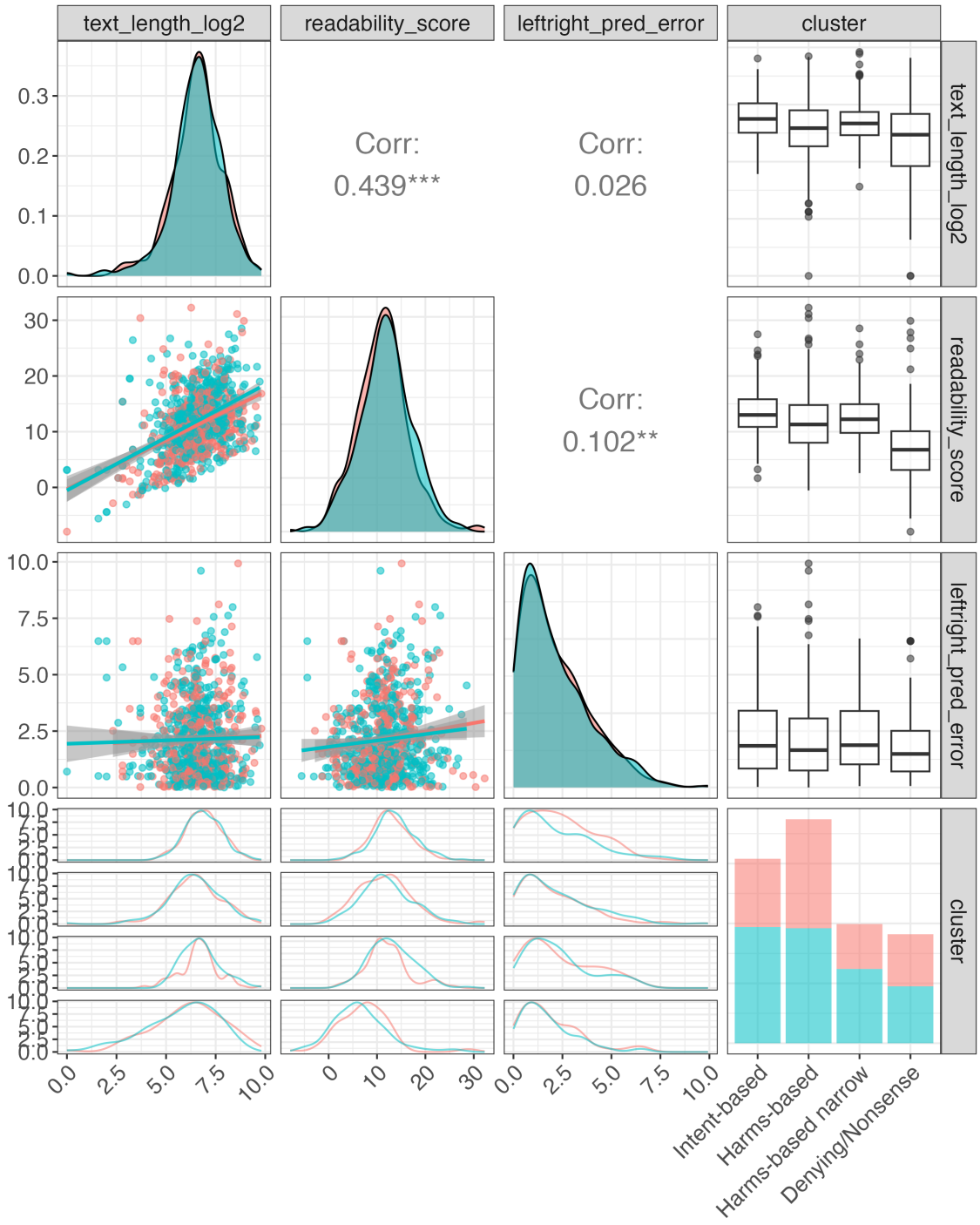


Figure 9: Indicator distributions and correlations, $n = 1,049$

Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on May 16, 2024 is identical to the printed version I submitted to the Examination Office on May 16, 2024.

DATE: 16.05.2024

NAME: Johanna Mehler

SIGNATURE: