Analyzing Hate Speech: A Brief Overview for Research Methods Classes

Emily M. Smith

Jeffrey M. Stanton

Yisi Sang

Syracuse University

Draft dated: April 18, 2023

Acknowledgement: This paper was developed with support from a small grant from the Association for Psychological Science

Analyzing Hate Speech: A Brief Overview for Research Methods Classes

Over the past decade, the integration of internet-based social media into the everyday lives of people has reached unprecedented levels (Saha et al., 2019). Yet just as the internet can be used for beneficial purposes, so can it be exploited to victimize people, through a variety of mechanisms including online bullying, doxxing, identity theft, and the widespread circulation of hate speech. While psychologists have studied the detrimental impact of hate speech on individuals, there are many open questions about both motivational aspects – why individuals choose to post and distribute hate speech – as well as the long term psychological impacts for those individuals and groups who are targeted (Saha et al., 2019). As such, psychologists are interested in new research methods that can provide insight into these issues.

One such method, text analysis, can provide unique insights into the occurrence and impacts of hate speech across online spaces. Powerful, yet easy-to-use techniques are now available for analyzing trends across large amounts of text. With the aid of machine learning techniques, social and behavioral scientists can identify meaningful patterns that could provide insights into the identities, mindsets, and motives of those who create and distribute online hate speech. Researchers can apply text mining and natural language processing techniques to process social media posts, develop multivariate models, and test theories about the psychology behind both posters and targets. This brief overview and introduction provides key context and background information for those interested in phenomena related to hate speech but who are unfamiliar with contemporary text analysis and processing techniques.

Definitions and Characteristics of Hate Speech

Opinions vary on what exactly hate speech is, how it can be defined, and how it can therefore be quantified in the research setting. After all, as is the challenge with language, its nature

is deeply impacted by a speaker's intent, the audience, and the context in which the language is used. Laaksonen et al. defines hate speech based on its capacity for damage and explains that hate speech "covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance" (Laaksonen et al., 2020, p. 3). This definition outlines that hate speech is characterized not necessarily by the nature of the language used, but the intended or actual impact. Hate speech may or may not contain language that describes violence, but the author often intends to incite violence – generally against targeted groups or against specific members of a group.

It is this emphasis on group membership which differentiates hate speech from bullying. Much like hate speech, bullying utilizes language whose effects can be seen as harsh, if not harmful to an intended audience. However bullying tends to be more generalized, targeting an individual's actions or identity as opposed to one's specific group membership. Bullying can consist of terms like "stupid" or "loser;" terms that fall under "generalized abuse, largely towards the appearance, interests, intelligence or previous posts of the recipient" (Infographic: Cyberbullying and Hate Speech Online, 2016). These terminologies call out a person's identity and can even be violent in nature. But, what distinguishes certain rhetoric as hate speech is that it is "abuse directed specifically toward a unique, non-controllable attribute of a group of people" (Infographic: Cyberbullying and Hate Speech Online, 2016). Pointing abuse towards an aspect of group membership, thus differentiating the victim as part of this scorned "outgroup," is what defines hate speech.

Another definition of hate speech, outlined by Saha et al. (2019), focuses on keyword identification, assigning greater power to specific language based on the ability of certain words to inflict damage. Insults, threats, slurs, as well as images can all qualify as hate speech, due to

their implicit or explicit intent to evoke violence (Saha et al., 2019). Under this definition, certain loaded language constructions distinguish hate speech from speech that is merely offensive. Saha et al.'s article includes a variety of terms which call specific attention to the receiver's "behavior, class, disability, ethnicity, gender, physical, race, religion, sexual orientation, and other" (Saha et al., 2019, p. 257). Consider the following example post: "dont assume your retarded worldview is correct" (Saha et al., 2019, p. 258). Although the post itself does not overtly call for violence upon the reader, the word "retarded" jabs at the reader's disability status. By Saha et al.'s (2019) definition, this jab, or loaded word, is hate speech as it brings with it a connotative assumption of violence.

These definitions provide us with two essential characteristics of hate speech: namely that hate speech is language that addresses members of a specific target group and which incites violence against that group. The working definition that this overview will use is thus that hate speech is language that implicitly or explicitly incites violence against an identifiable demographic group.

Linguistic Qualities of Hate Speech

Hate speech takes on many forms, such as messages that overtly call for harm, or insulting words or phrases, often referred to as slurs. One such form is covert hate speech, which flourishes in online settings. As the name suggests, covert hate speech includes messages "that are not necessarily direct expressions of hate but support the circulation of hate and are used to stir up hatred as well as support hate communities" (Laaksonen, 2020, p. 3-4). Covert hate speech is often hyper-effective against its targets and difficult for others to recognize, given the fact that it is encoded as everyday speech.

In a similar vein, hate speech can also come in the form of dog whistles; bits of rhetoric that, while seemingly innocuous, carry with them connotations for hate and violence. Dog whistles often signify allegiance to an ingroup as well as animosity towards an outgroup. Dog whistles do not always read to a general audience as inherently hateful, and may pass as harmless or nonsensical bits of rhetoric (Bhat and Klein, 2020). Due to the supposed harmlessness, this language can sometimes be circulated without negative repercussions for the sender.

Impacts of Internet Distribution

Internet-based social media platforms have proven to serve as hotbeds for hate speech as their features allow for efficient sharing of ideas. Even on mainstream platforms with extensive content moderation, hate speech has adapted to fly under the radar, leaning on coded content to convey meaning without triggering moderation processes. Furthermore, the internet's proclivity for anonymity can impact the ability to assess identity context. As suggested by Almagro et al. (2022), identity context is a good identifier of hate speech; however, on most online platforms, users can maintain anonymity by hiding behind screen names, and otherwise failing to disclose information that could allude to related attitudes. Thus hate speech flourishes in an online environment since "under the veil of (semi)-anonymity…perpetrators receive reinforcement from like-minded haters, making hatred seem normal and acceptable" (Saha et al., 2019, p. 256).

Psychological Impacts of Hate Speech

Hate speech contains a wide capacity for damage, regardless of whether particular speech examples explicitly incite physical violence. In fact, much research suggests that the impact of speech found on online platforms is largely psychological (Barendt, 2019; Cowan & Hodge, 1996; Saha et al., 2019).

Impacts on Individuals Targeted by Hate Speech

Research suggests that victims exposed to hate speech "[show] similar three-phase consequences of feelings (affect), thoughts (cognition), and actions (behavior) as when experiencing other traumatic events. Subsequently, the receivers of hateful speech may experience psychological symptoms, similar to post-traumatic stress disorder" (Saha et al., 2019, p. 256). Typical symptoms manifest as heightened levels of stress and anxiety. Victims of hate speech may also experience a decreased sense of self-worth and lowered sense of enjoyment regarding social activities. This withdrawal from activities is one of the behaviors most indicative of hate speech exposure as "prolonged experiences of discrimination may also lead to increased vigilance and shape expectations of minority group members regarding possible future discriminatory situations" (Wypych and Bilewicz, 2022, p. 2). This hypervigilance is perhaps one of the most insidious side effects of hate speech exposure as it causes victims to further isolate themselves from positive social circles, feeding into the cycle of self-isolation and depression. This depression can manifest itself in a variety of behavioral changes; a decline in academic performance, work function, or even a generalized malaise (Saha et al., 2019). Researchers should be vigilant about noting the impacts of hate speech from both a psychological and a behavioral standpoint.

Impacts on Individuals Who Post Hate Speech

Just as exposure to hate speech can come at the detriment to the receiver, so can the usage of hate speech come at the detriment to the sender or "poster." While it may seem counterintuitive to view the perpetrators of hate speech as fellow victims of this rhetoric, the hostility expressed by hate speech often comes in response to a perception of threat. Miller (2021) posits that hate speech is tied to a skewed sense of morality as perpetrators act out of a desire to protect themselves from the imagined threat of the outgroup. As such, speakers may find themselves with those same

symptoms characteristic of those victimized by hate speech, including hypervigilance. Differences emerge, however, with respect to subsequent behavioral changes. While victims of hate speech may self-isolate, speakers often flock to their peers, becoming more deeply embedded into an ingroup identity. Within groups of like-minded individuals they "feed each other's visions of the world and anger towards the outgroup" (Miller, 2021). These visions of anger also have the capacity to incite violence against members of the outgroup, so while the impacts of hate speech are primarily psychological, there is always potential for physical harm, especially when speakers fall into paths of extremism.

Extremism is initiated through the process of radicalization. The example Miller provides is that of the #stopthesteal movement, in which posts under this hashtag resulted in a far-right echo chamber whose users then went on to combat the perception of a threat in the form of the January 6th, 2021 insurrection. The leap from hashtag usage to tangible action can be best understood through the steps of radicalization namely "(1) arousal of the goal of significance… (2) identification of terrorism/violence as the appropriate means to significance, (3) commitment shift to the goal of significance and away from other motivational concerns resulting in that goal's dominance" (Kruglanski et al., 2014, p. 74). Already immersed in a tense political climate, Twitter users recognized a "goal of significance" - a desire to preserve American democracy from the perceived threat of the outgroup. Regardless of the validity of said threat, ingroup user's fears were reaffirmed by the voices in the #stopthesteal echo chamber. From there a goal, or "means of significance" was created, ushering users into steps two and three in which they were actively combatting a perceived threat through acts of tangible violence. Though the use of the hashtag was not inherently violent, it was the ideologies promoted under it that made this rhetoric hateful, and opened a door into realms of political extremism.

Identifying and Categorizing Hate Speech

In order to conduct research on hate speech, it is often necessary to have methods of identifying examples of hate speech and classifying or scaling them in ways that facilitate analysis or appropriate study design. When setting out to identify hate speech, researchers have a myriad of methods at their disposal to comb through online content and extract examples of hateful rhetoric. One such method, keyword identification, can prove useful when looking for examples of hate speech that utilize the set phrases, slurs, and dog whistles as described above. As these instances of language are pre-established, they can be recognized analytically and coded, classified, or scaled as appropriate.

For keyword identification to be done effectively, researchers much first establish a lexicon from which they are pulling the terms (Saha et al., 2019). Development of such lexicons usually relies on human judgments about hateful rhetoric. One example of this type of database is known as *Hurtlex* and is discussed in more detail in a subsequent section. There is no single universally accepted lexicon, and indeed, this would be impractical due to differences in community and cultural standards and because of the rapid evolution of linguistic usage.

Some social media platforms use keyword identification as one (or the only) basis of their content moderation methods. Due to user guidelines specific to each social media platform, there are different standards of acceptability in terms of what vocabulary is flagged and removed from a site. Those posting hate speech often develop familiarity with these guidelines, as part of efforts to circumvent them. Efforts to avoid content moderation may take the form of symbols, altered phrasings, or even purposeful misspellings, making the maintenance of a lexicon time-consuming and expensive (ElSherief et al., 2018)

Larger social media companies have developed more complex, automated approaches to the detection of hate speech. Machine learning models can overcome the limits of using a fixed lexicon by detecting more subtle linguistic patterns that are present (or absent) in hate speech. These models work by taking "samples of labeled text to produce a classifier that is able to detect the hate speech based on labels annotated by content reviewers" (MacAvaney et al., 2019, p. 7). Machine learning systems can also be fine-tuned via additional training, for example by learning new patterns based on future decisions of human content moderators.

Limitations of Linguistics: Understanding the Context of Hate Speech

From a research perspective all methods that focus on linguistic patterns – either from a lexicon or a machine learning system – face an important limitation with respect to the context that surrounds the creation and distribution of hate speech (MacAvaney et al., 2019). Speaker identity is an important factor in the connotative meaning of certain words and phrases, yet this is something a purely linguistic method would be unable to recognize (MacAvaney et al., 2019). Likewise, methods of examining hate speech text generally lack the ability to recognize power structures. For example, oftentimes victimized groups will respond to hate speech with words of aggression. While this rhetoric lacks the intention of violence that is indicative of hate speech, the usage of certain words or phrases would have these posts flagged as such (MacAvaney et al., 2019). A poster's social standing may also define what capacity for violence they possess. The language utilized by a politician may be loaded with additional connotations due to that person's ability to evoke political action. This distinction is not one that could be easily captured by a machine learning model. Furthermore, some language that seems violent may not be hate speech, whereas seemingly innocuous phrases may carry with them an immense capacity to evoke violence and harm.

Gathering and Analyzing Research Data

With some understanding of the sources, language, and context of hate speech, researchers can develop valuable insights by gathering and analyzing research data. One ready source arises through the harvesting of publicly-available available internet data. Though atypical in psychology, this methodology for "archival" data gathering can provide unique insights into how hate speech impacts online spaces and the people who create or are targeted by hate speech. One example of a method for harvesting online speech is known as VOSONsml (Ackland et al., 2006). This software integrates with the R analytical system and provides methods to connect with and extract data from three large social media platforms.

After obtaining data containing possible examples of hate speech, researchers can use a variety of pre-established linguistic databases that identify hate speech based on specific keywords and phrases. For example, Bassignana et al. (2018) published a multilingual lexicon of offensive speech examples known as Hurtlex. An actively maintained project, "The Weaponized Word" has a lexicon of approximately 7500 hate-related words and phrases. Researcher Bruno Martins maintains four lexicons of different types of hate speech in English, Spanish, and Portuguese on the Github repository platform. Additionally, many such databases are available to academic researchers at no cost. Overall, there is a wide array of databases, and other textual resources made available for researchers. A table provided in the appendix outlines key attributes of several examples of such lexicons. When examining online content researchers can compare chosen selections to that of the database to verify the occurrence of hate speech (MacAvaney et al., 2019).

When building a lexicon, it is appropriate to work with a variety of datasets in order to ensure a broader scope of available terminologies for the researcher to identify. Due to the ever-evolving nature of hate speech, variations in spellings and usages exist for different terms –

differences that may not be recognized by one dataset alone. By using multiple datasets researchers may equip themselves with a more holistic knowledge of terms, and perform more precise identification (Chiril et al., 2021). For example, some lexicons may focus on hate speech in a particular language, and thus offer a more specific and nuanced look at the sort of cultural and psychological relations that occur within that context. Conversely, other datasets may pull from a variety of languages, offering up a broader pool of terms. Datasets with added emphasis on visual rhetoric may also prove helpful in creating a well-rounded lexicon. For research purposes, the essential characteristic in devising a lexicon is that it pulls from a diverse array of data.

A more complex approach can be conducted with the aid of a variety of machine learning models including deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The use of deep learning models depends very much on the quality of training data; thus researchers may need to try several sources of training data before settling on the development of a model. A recent study conducted by ElSherif et al. (2018) utilized cross-model comparisons. Researchers used a psycholinguistic lexicon software, LIWC2015, to measure specific psychological dimensions of online text, as well as a mixed-effect topic model, SAGE, "to analyze the salient words for each category of hate speech keywords" (ElSherif et al., 2018, p. 45). Indeed, these were just two of many models used by researchers in the study to measure instances of hate speech across a Twitter thread. While the parameters of one's research may not be identical to that of this example, the described methodology does provide best practice as to how one should go about harvesting their data. Parallel and multiple usage of models as well as cross-model comparison gives researchers a larger, more holistic dataset from which they can begin to make an analysis.

In the context of hate speech research, text analytics can go beyond simply identifying instances of hate speech and instead work to discern larger trends, such as the poster's identity or intent. With text analytics, "emphasis is on inferring information that may be implicitly conveyed by the text. A classic example is to infer the sentiment, whether positive, negative, or neutral, of a text or its components" (Srinivasan, 2020, p. 180). Often obtained with the use of machine learning models, sentiment analysis allows researchers to score phrases based on certain connotative meanings. While sentiment analysis is often used by marketing teams as a means for combing customer feedback, the applications in psychology support connecting the content of speech examples with other variables. This use of text analytics has been employed in the past "to infer latent demographic attributes such as gender, race, age group, and location from an individual's posts… and the person's online behavior" (Srinivasan, 2020, p. 181). However, these machine learning models can go deeper, and also assess "social, health, and psychological attributes, such as personality dimensions, well-being, political leanings, health status, sexual preferences, religiosity, and so on" (Srinivasan, 2020, p. 181). In the context of hate speech, this provides the researcher with insights into a poster's mental health status, propensity for violence, or affiliation with specific extremist groups.

Prevalence Data

The use of text harvesting and machine learning models can also aid researchers in gathering qualitative data pertaining to hate groups. Large amounts of online content can be feasibly combed through by trained machine learning models and assessed for poster intent and action. This kind of tracking works by having "built-in machine learning capabilities... semi-automatically classify and extract sentiment from posts," based on fine-tuning via a training dataset which aids in the model's ability to collect and then interpret its data (UN Global Pulse and

UNHCR Innovation Service, 2017, p. 6). With training, the models learn to recognize hate speech from a set of linguistic parameters, soon recognizing other factors such as locale, time, and trends of communication. As a result, larger trends can be ascertained from data, potentially providing researchers insight into what causes spikes in hate group activity, hate speech, or even yielding information on what posters are more susceptible to paths of extremism.

Geographical comparisons can also be made using the data gained from text analysis. Despite this speech populating online spaces, "hate crimes as well as hate groups have a spatial component… hence, to understand the potential relationship between hate groups and hate crimes, their patterns of spatial correlation and dependence must be considered" (Jendryke and McClure, 2019, p. 2-3). As such text analytics can provide useful insights into the geographical characteristics of hate speech, such as tracking slang back to specific regions or tracing the movements of online posters in accordance with the information pulled from their content. As such, researchers can train machine learning models to identify activity withing certain locales and apply the use of certain identifiers to the real-time posts of specific hate groups.

Another application of this research is time-based analysis, which operates by utilizing time-stamped data gathered from text harvesting to provide a picture of how hate groups change over a particular period. While time-based analysis is not new to the field of psychology, this style of analysis, when facilitated by text-harvesting data, can give greater insight into certain hate groups. In the context of hate speech research, "online behaviors can now be quantified and tracked in real-time, leading to an accessible and rich source of time series data" (Jebb et al., 2015, p. 1). Through the use of machine learning, this predictive time-based analysis can be applied to online content, granting researchers access to patterns in posting, event correlations, and other trends.

Subjective Reports

Psychologists are generally well versed on self-report methods, including self-reports provided by victims of hate speech. Subjective reporting remains a good way of gathering data on the psychological impacts of hate speech. One such approach would be speaking to victims directly, either via an in-person assessment or with the aid of an online platform. A researcher may wish to hold individual interviews or conduct a focus group, in which questions are posed to a group of participants, yielding a dialogue that may give greater insights into the research topic. Survey research can also be conducted by eliciting respondents from online spaces (Jhangiani et al, 2020). This technique carries ethical implications, so researchers must be intentional with how they distribute their surveys, and what sort of questions they include. Additionally, researchers should seek out diverse samples because particular groups "might have different norms with regard to the acceptability of such behaviors…they might also have different networks of social support, which could affect how they cope with experiences of hate speech" thus skewing the data (Wypych and Bilewicz, 2022, p. 7).

Lastly, and perhaps most beneficial, is participant observation. This method requires that "researchers become active participants in the group or situation they are studying…the basic rationale for participant observation is that there may be important information that is only accessible to, or can be interpreted only by, someone who is an active participant in the group or situation" (Jhangiani et al., 2020, p. 170). Given that the data collected will be data pulled from online sites and forums, entry into said spaces can help researchers better familiarize themselves with situations where hateful rhetoric is disseminated. The research ethics related to researcher involvement in active online forums are complex, so studies using this technique should be thoroughly vetted by institutional review boards.

Research Ethics Related to Text Harvesting and Analysis

When conducting this research, it is imperative that researchers familiarize themselves with the ethical concerns that surround text harvesting. Just as the examination of in-person behavior falls under strict ethical scrutiny, so do the observations gathered when behaviors are examined in an online setting. Researchers should proceed with caution when gathering data from online spaces because of dilemmas related to user consent. Overall, "it can be very difficult to gain consent for the research use of harvested data, particularly because contact information may not be available, especially if the data are harvested sometime after they were produced" (Zia et al., 2020, p.3). The issue of privacy and consent becomes critical when researchers expect to publish direct quotes from these online forums. While online spaces may appear to be public settings with no expectation of privacy, the assumptions of actual users may differ which can cause the possibility of perceived and actual harms related to a researcher's work.

Regarding data gathered from online settings, "it has been suggested that each type of online research method (i.e., observational, interactive, or survey/interview research) is highly contextual" (Gupta, 2017, p. 5) and therefore expectations for direct engagement with researchers will vary among participants. The standards for consent become murkier when data is gained from text harvesting. In these instances, there may still be ways for researchers to maintain good ethical practices. For example, researchers can seek consent from forum moderators who can serve as a proxy for participant consent (Zia et al., 2020). Importantly, in most instances, pseudonyms or randomly generated identification numbers should be used in place of names or usernames (Zia et al., 2020). Although these methods are not perfect, they are likely to be helpful in preserving participant privacy.

Compassion Fatigue

Ethical concerns are not only posed in relation to those giving and receiving hate speech but can also impact those who study it as well. Researchers working in this area often desire to learn from and aid those who are victimized by hateful rhetoric. Yet this compassion for victims may pose harm to the researcher. Compassion fatigue is a term often used regarding social work which refers to a heightened stress reaction brought on by prolonged exposure to the hardships of others, one whose effects can drastically diminish a sufferer's health, well-being, and ability to feel empathy (Paiva-Salisbury and Schwanz, 2022). A combination of burnout and proximity to trauma characterizes compassion fatigue. Unfortunately, "this places certain occupations, such as healthcare, emergency and community service workers, at an increased risk" (Cocker and Joss, 2016, p. 2). Just as proximity to hate speech can cause adverse effects on the sender's target, so can it negatively impact researchers. This wide net of impact is characteristic of online hate speech as "it does not have to be directed to an individual, but it can affect any person who encounters such an utterance" (Wypych and Bilewicz, 2022, p.1). Merely encountering hate speech can be enough to harm an individual. As such, researchers wishing to pursue this line of work should proceed with the knowledge of the sort of strain their work entails.

One of the first steps in addressing compassion fatigue is learning to recognize and monitor symptoms of anxiety, stress, fatigue, and apathy (Paiva-Salisbury and Schwanz, 2022). Early recognition brings the opportunity for early intervention. Furthermore, routine self-care practices should be advocated for. Inclusion of self-care training into psychological education is imperative because it "is a critical component of CF resilience because it can ameliorate the harmful effects of trauma work" therefore its (Paiva-Salisbury and Schwanz, 2022, p. 40). Self-care can be

obtained in formal settings, such as wellness seminars, or instructional classes, as well as in informal settings such as peer support groups.

Conclusion

Text harvesting offers many opportunities for exploring issues related to online hate speech, such as the devastating effects such rhetoric can have on its recipients. Using simple keyword identification techniques or more complex machine learning models, researchers can comb through online posts, isolate instances of hate speech, and analyze them for useful insights. Using text-based methods in conjunction with more traditional quantitative research techniques, researchers can explore how hate speech impacts the brain and behaviors, as well as how this phenomenon perpetuates hate, violence, and extremism in contemporary societies.

References

Ackland, R., O'Neil, M., Standish, R. K., & Buchhorn, M. (2006). VOSON: A Web services approach for facilitating research into online networks.

Almagro, M., Hannikainen, I. R., & Villanueva, N. (2022). Whose Words Hurt? Contextual Determinants of Offensive Speech. *Personality and Social Psychology Bulletin*. 48(6). 937-953. https://doi-org.libezproxy2.syr.edu/10.1177/01461672211026128

Barendt, E. (2019). What is the harm of hate speech?. *Ethical Theory and Moral Practice*, *22*, 539-553.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018* (Vol. 2253, pp. 1-6). CEUR-WS. https://iris.unito.it/handle/2318/1684807

Bhat, P., Klein, O. (2020). Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In: Bouvier, G., Rosenbaum, J. (eds) *Twitter, the Public Sphere, and the Chaos of Online Deliberation*. (pp. 151-172). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-41421-4_7

Chau, M. and Xu, J. (2008). Using web mining and social network analysis to study the emergence of cyber communities in blogs. In H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor (Eds.) *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*. pp. 473-494. Springer. https://doi.org/10.1007/978-0-387-71613-8

Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., Patti, V. (2021). Emotionally informed hate speech detection: A multi-target perspective. Cognitive Computation, 14(1), 322–352. https://doi.org/10.1007/s12559-021-09862-5

Cocker, F., & Joss, N. (2016). Compassion Fatigue among Healthcare, Emergency and Community Service Workers: A Systematic Review. *International journal of environmental research and public health*, *13*(6), 618. https://doi.org/10.3390/ijerph13060618

Cowan, G., & Hodge, C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, *26*(4), 355-374.

ElSherief, M., Kulkarni, V., Nguyen, D., Yang Wang, W., & Belding, E. (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15041

Gupta, S. (2017). Ethical issues in designing internet-based research: Recommendations for good practice. *Journal of Research Practice*, *13*(2), Article D1. Retrieved from http://jrp.icaap.org/index.php/jrp/article/view/576/476

Hu, Y. and Adams, B. (2021). Harvesting Big Geospatial Data from Natural Language Texts. In M. Werner and Y. Chiang (Eds.) *Handbook of Big Geospatial Data*. pp. 487-507. Springer. https://doi.org/10.1007/978-3-030-55462-0

"Infographic: Cyberbullying and Hate Speech Online." Brandwatch. Ditch the Label, 2016. https://www.brandwatch.com/reports/cyberbullying-2016/.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology*, *6*, 727. https://doi.org/10.3389/fpsyg.2015.00727

Jendryke, M. and McClure, S. C. (2019). Mapping crime – Hate crimes and hate groups in the USA: A spatial analysis T with gridded data. *Applied Geography*, 111, 1–8. https://doi.org/1016/j.apgeog.2019.102072

Jhangiani, R. S., Chiang, I. A., Cuttler, C., and Leighton, D. C. (2020) *Research Methods in Psychology.* (4th ed.) Kwantlen Polytechnic University. https://kpu.pressbooks.pub/psychmethods4e/

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The Psychology of Radicalization and Deradicalization: How Significance Quest Impacts Violent Extremism. *Political Psychology*, *35*, 69–93. http://www.jstor.org/stable/43783789

Laaksonen, S.-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., Pöyhtäri, R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. Frontiers in Big Data, 3. https://doi.org/10.3389/fdata.2020.00003

MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: Challenges and solutions. PLoS ONE 14 (8): e0221152. https://doi.org/10.1371/journal.pone.0221152

Miller, J. (2021, December 16). Hate speech and online extremism focus of USC Study. *USC News*. Retrieved September 29, 2022, from https://news.usc.edu/195881/online-extremism-linked-to-shared-moral-beliefs/

Paiva-Salisbury, M. L., & Schwanz, K. A. (2022). Building Compassion Fatigue Resilience: Awareness, Prevention, and Intervention for Pre-Professionals and Current Practitioners. *Journal of health service psychology*, *48*(1), 39–46. https://doi.org/10.1007/s42843-022-00054-9

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and Psychological Effects of Hateful Speech in Online College Communities. *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference*, *2019*, 255–264. https://doi.org/10.1145/3292522.3326032

Sang, Y. (2021). Understanding And Incorporating Subjective Human Judgments In Social Computing Classifiers: A Case Study Of Hate Speech Detection (dissertation).

Srinivasan, P. (2020). Text Mining: A Field of Opportunities. In S. E. Woo, L. Tay, and R. W. Proctor (Eds.), *Big Data in Psychological Research*. pp. 179-200. American Psychological Association. https://www.jstor.org/stable/j.ctv1chs5jz.12

UN Global Pulse and UNHCR Innovation Service. (2017, September). Social Media and Forced Displacement: Big Data Analytics & Machine-Learning. [White paper]. The UN Refugee Agency. Retrieved from https://www.unglobalpulse.org/document/social-media-and-forced-displacement-big-data-analytics-and-machine-learning/

Woo, S. E., Tay, L., & Proctor, R. W. (Eds.). (2020). *Big Data in Psychological Research*. American Psychological Association. http://www.jstor.org/stable/j.ctv1chs5jz

Wypych, M., & Bilewicz, M. (2022 ,January 31). Psychological Toll of Hate Speech: The Role of Acculturation Stress in the Effects of Exposure to Ethnic Slurs on Mental Health Among Ukrainian Immigrants in Poland. Cultural. *Diversity and Ethnic Minority Psychology*. Advance online publication. http://dx.doi.org/10.1037/cdp0000522

Zia, S., De Lancey, C., Regan, P., & Burkell, J. . (2020). There for the Reaping: The Ethics of Harvesting Online Data for Research Purposes. *Proceedings of the Annual Conference of CAIS / Actes Du congrès Annuel De l'ACSI*. https://doi.org/10.29173/cais1151

Appendix

No.	Name of the Resource	Brief Description	Links where it can be found	Size of the Database	Number of records/instances	Description of the main columns/variables in the database
1	Hate Speech and Offensive Language Dataset	Data is from Twitter wherein different users use hate speech/offensive languages	https://www.kaggle.com/mrmorj/hate-speech-and-offensive-language-dataset	2.55MB	25296	hate_speech, offensive_language, neither - users who judged the tweet count - number of CrowdFlower users who coded each tweet class - class label for majority of CF users
2	Hate Speech Dataset from a White Supremacy Forum	This data is from a supremacy forum where the users have posted their data	https://github.com/Vicomtech/hate-speech-dataset/blob/master/annotations_metadata.csv https://paperswithcode.com/paper/hate-speech-dataset-from-a-white-supremacy	341 KB	10945	file_id -> the forum file id user_id -> to identify the user subforum_id label -> data label for hate/nohate identification
3	ETHOS Hate Speech Dataset	This repository contains a dataset for hate speech detection on social media platforms, called Ethos.	https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset	62 KB	433	violence (if it incites (1) or not (0) violence), directed_vs_general (if it is directed to a person (1) or a group (0)), and 6 labels about the category of hate speech like, gender, race, national_origin, disability, religion and sexual_orientation
4	A Benchmark Dataset for Learning to Intervene in Online Hate Speech	These datasets have been taken from Gab and Reddit	https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech	8.73 MB & 7.2 MB	5024	id -> the ids of the post in a conversation segment text -> the text of the posts in a conversation segment hate_speech_idx -> a list of the indexes of the hateful posts in this conversation response -> a list of human-written responses

5	Multilingual and Multi-Aspect Hate Speech Analysis	This dataset contains tweets and annotations in different languages	https://github.com/HKUST-KnowComp/MLMA_hate_speech	710 KB	5647	the hostility type (column: tweet sentiment), hostility directness (column: directness), target attribute (column: target), target group (column: group), annotator's sentiment (column: annotator sentiment)
6	Hate Speech and Offensive Content Identification in Indo-European Languages	The data is collected from Facebook and Twitter from the Indo-European region	https://hasocfire.github.io/hasoc/2019/dataset.html	1.16MB	5852	text_id -> the id of the text text -> the actual text that contains the offensive data task_1 -> hate/offensive/neither task_2 -> hatespeech/offensive/profane task_3 -> targeted/untargeted
7	Detecting cyberbullying in online communities (World of Warcraft)	The data is collected from Twitter	http://ub-web.de/research/	380KB	16146	tweetid -> the id of the tweet bullyLabel -> 0 neutral/ 1 harrasment labelUser
8	A Quality Type-aware Annotated Corpus and Lexicon for Harassment Research	Twitter data. Details of the task: Racism, Sexism, Appearance-related, Intellectual, Political	https://github.com/Mrezvan94/Harassment-Corpus	9 KB	453	The corpus is divided into six categories: 1) Sexual 2) Appearance-related 3) Intellectual 4) Political 5) Racial 6) Combined