Evasive offenses: Linguistic limits to the detection of hate speech

Baden, Christian

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Baden, C. (2023). Evasive offenses: Linguistic limits to the detection of hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 319-332). Berlin https://doi.org/10.48541/dcr.v12.19

Digital
Communication
Research.de

Recommended citation: Baden, C. (2023). Evasive offenses: Linguistic limits to the detection of hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 319–332). Digital Communication Research. https://doi.org/10.48541/dcr.v12.19

Abstract: As long as we have attempted to sanction untoward speech, others have devised strategies for expressing themselves while dodging such sanctions. In this intervention, I review the arms race between technological filters designed to curb hate speech, and evasive language practices designed to avoid detection by these filters. I argue that, following important advances in the detection of relatively overt uses of hate speech, further advances will need to address hate speech that relies on culturally or situationally available context knowledge and linguistic ambiguities to convey its intended offenses. Resolving such forms of hate speech not only poses increasingly unreasonable demands on available data and technologies, but does so for limited, uncertain gains, as many evasive uses of language effectively defy unique valid classification.

DOI 10.48541/dcr.v12.19

Christian Baden

Evasive Offenses

Linguistic limits to the detection of hate speech

1 Introduction

In an arms race, the offender is typically one step ahead: As defensive technologies are largely designed to fend off known threats, new offensive strategies continue to challenge the development of ever more sophisticated responses. Some threats remain durably beyond the reach of an effective defense, either because they are too unpredictable, or because suitable defenses would infringe in unjustifiable ways upon the liberties of those that they purport to defend, and it is preferable to tolerate the remaining risk. In this intervention, I will argue that this is true not only in security, in cybersecurity, and many other domains, but also in the detection of hate speech.

In the following, I will sketch a rough, but I hope informative caricature of the arms race that has unfolded over the past decades between hate speech and opposing efforts at maintaining civil discourse in online environments. Specifically, I will point out major advances in available technology, as well as specific evasive strategies adopted by users of hate speech (or other sanctioned language uses) in an effort to elude these technological filters. As I will show, many earlier technological advances have successively improved our capacity to detect hate

speech, but have focused on its comparatively plain variants—notably, misspellings, neologisms, and polysemic expressions. With the progressing deployment of context-enriched, AI-based filtering (Kumaresan & Vidanage, 2019), those uses of hate speech that continue to evade unique classification increasingly rely on cultural and situational context knowledge as well as linguistic ambiguity to convey intended offenses. Resolving such uses of evasive language not only poses demands on available data and language processing technologies that quickly exceed defensible dimensions; in many cases, it may even prove impossible to obtain a unique, valid classification. To the extent that further gains are increasingly unlikely, incurring sensitive biases and raising serious ethical objections, we might as well acknowledge that hate speech ultimately constitutes a *social* problem—one that may well be contained, but cannot be resolved, by technical means.

2 The evasive nature of language

As a starting point, we need to acknowledge that language lives, in a sense that does not stretch the metaphor very far (Mufwene, 2001): Words and meanings evolve to match new realities and address new purposes, and language uses respond to the socio-cultural and socio-technical environments that they inhabit. Where it is challenged, language adapts and finds new ways to meet its purpose—for it is the purposes, not the words, that ultimately govern how language is used. Accordingly, any effort to sanction specific uses of language provokes opposing efforts to achieve the same objective while circumventing the sanction (e.g., Gerrard, 2018).

While this is true generally for how language is used, it is particularly true for what linguists call speech acts (Searle, 1969), that is, the use of language not merely to describe, express or otherwise inform, but to elicit certain social effects. Given that this 'pragmatic' use of language for managing social relations is inherently controversial, all languages have developed manifold strategies for committing the same speech act, using different words and expressions depending on its sanctioning in a social context. In circumstances where we don't (have to) fear sanctioning, we may say *in plain words* what we mean ("What you say is absurd"), but for each use, there is typically a whole bouquet of expressions that convey the same meaning in ways but are more likely to pass as acceptable in

situations governed by more restrictive behavioral rules (e.g., "I seriously doubt that," "Oh please, let's not go there again," "Right" [sarcastically]; Bavelas et al., 1990). Using various forms of evasive language uses, we can criticize our partner's cooking (e.g., "Interesting..."), express our disdain for our boss (e.g., "Our wonderful leader"), offer a bribe ("I am sure we can find an agreeable solution...") or inquire whether someone might be interested in sexual relations (e.g., "Want to come up for one more drink?") – all the while maintaining a plausible pretense that this was not our intended meaning, should the response be adverse (Gruber, 1993; Obeng, 1997). As long as we could get ourselves into trouble by what we say, evasive language uses have been there for us to dodge expected sanctions. Accordingly, when algorithmic sanctioning entered the stage of digital communication, language was ready for it.

3 The words that weren't so

As more or less anything in natural language processing technologies, also the sanctioning of inappropriate speech started as a list of keywords—typically, of more or less openly derogatory labels or references to racist, anti-Semitic, misogynist or otherwise hostile discourses (Zelenkauskaite et al., 2021). Noticing that certain terms were suppressed, users of early chat rooms and forums quickly learned to use creative spellings, truncated words (a particularly interesting case is "f***"/"f-ing," where written—and to some extent even spoken—language use redacts itself in anticipation of being redacted, thereby evading redaction while simultaneously marking the sanctioning of the expressed meaning; Fairman, 2006), and acronyms. Leet (the replacement of certain letters by numbers) was one outcome, and many para-linguistic symbols (e.g., the "(((They)))" meme; Tuters & Hagen, 2020) and neologisms (e.g., "cuck," "libtard"; Hodge & Hallgrimsdottir, 2020) were born to outsmart the filtering algorithm. Keyword lists evolved and grew in pace, trying to catch any known and increasingly conventionalized spellings, and fuzzy matches increasingly enabled algorithms to also catch simple variations, such as (accidental or deliberate) misspellings and leet.

At the same time, the redaction of any expressions used as swearwords, inappropriate comments or hate speech rapidly revealed an important limitation of such keyword-based strategies, which chiefly derived from two main problems.

On the one hand, redacting or posts containing certain words effectively disabled also discussions wherein offensive terms were used not to trade insults, but to negotiate communication norms and their policing (e.g., debates about real or hypothetical uses of offensive terms in other communication environments). On the other hand, problems arose when words were used to convey offensive meaning that also had other uses (Magu et al., 2017).

4 The words that weren't that

In response to the possibility to use potentially offensive terms in non-offensive ways, one key strategy was to augment existing keyword lists with additional disambiguation criteria. For instance, algorithms might distinguish whether a term was used as part of a quote, thus enabling users to quote and criticize the others' words or negotiate communication norms without triggering a sanction. Longer expressions could be considered to distinguish between "white trash," "white trash can," and "this white trash can lick my..." (Warner & Hirschberg, 2012). Algorithms could be taught to distinguish uses of "swine" within and outside an agricultural context, or recognize the token "Fucking" as a reference to the so-named town in Austria (recently renamed Fugging).[1] Of course, any such rule-based filters could easily be gamed, as users figured out which combinations the algorithm might catch or tolerate, generating new expressions and linguistic obfuscations that were plain to the reader, but unclear to the machine. Still, context-based disambiguation constituted an important advance in the detection of hateful speech.

That said, disambiguation needs by far outstretched the capacity of text-based algorithms. One problem arises from the use of terms that are mostly used in benign ways (e.g., "chocolate," "snowflake," "Skype"; Magu et al., 2017) but can be also used to express contempt and hatred (e.g., as racial slur). As the specific meaning of such terms often arises from the wider context of a statement, valid disambiguation rules are near-impossible to define. Moreover, especially group labels such as "gay" or "Jew" can be used in both offensive and benign ways in more or less identical linguistic contexts (e.g., "seems everyone is gay there"),

1 https://www.politico.eu/article/austrian-village-of-f-king-to-be-renamed-fugging/

while the meaning depends on who is saying these words, and to whom: their derogatory potential rest half in the inaccuracy of their use (e.g., calling a man "little girl"; Schmidt & Wiegand, 2017), and half in the subcultural valuation of their denoted meaning (e.g., among anti-Semites, homophobes; Hodge & Hallgrimmsdottir, 2020). Not only do subcultures develop their own, idiosyncratic vocabularies and expressions to express hostility in oblique, identity-coded ways, multiplying the range of indicators and rules require consideration (e.g., in German youth culture, "victim" can denote a contemptible weakling and fool; in the misogynic Incel [involuntary celibate] movement, "Stacy" constitutes a sexually objectifying, resentful reference to a pretty woman; Jaki et al., 2019); but the very same expression can often be read to convey or not convey an insult, depending on the reader's habitual language use and awareness of communication contexts (see Litvinenko in this volume, for the various layers of such contexts).

Moreover, ethical issues arise from defining membership categories such as "Jew," "gay," "feminist" or "black" as potentially offensive terms, and any mistaken suppression of such references may justly raise public outcry.

5 The words that weren't needed

With the advance of machine learning based natural language processing, filters once again appeared to catch up with the manifold variations in language use in context. Relying on an appraisal of entire textual contributions and large databases of reference cases to distinguish textually similar, but semantically or pragmatically different language uses, supervised algorithms are capable of flagging problematic uses with much improved nuance and accuracy (Schmidt & Wiegand, 2017). Still, blind spots exist wherever relevant terms are absent in the reference corpus, or if there are too few reference cases to draw confident inferences. While the problem arises primarily for rare expression and can be mitigated by more inclusive training samples, this strategy quickly becomes unwieldy for highly heterogeneous communication contexts, where very many different uses may require consideration. Especially considering the reliance of machine learning algorithms on past language use, the constant evolution of digital discourse continuously weakens the predictive power of past reference cases. New events and situations enable new variations in the use of suspect words that

the algorithm could impossibly predict, and language users continually develop new ways to express their contempt. Moreover, machine learning strategies are particularly slow to adapt to new or rare language uses, as they depend on a sufficient number of cases to be manually rated, included in the training data, and accumulated to sustain confident algorithmic disambiguation.

Beyond those challenges raised by terms that may or may not express contempt, yet greater challenges arise from the expression of contempt without resort to potentially offensive terms. As machine learning tools tend to err in the direction of terms' more common usages, they are unlikely to recognize hate speech conveyed by means of entirely innocuous words (e.g., "back in the day, we would have put them on a train to the East," here conveying a veiled holocaust reference). Based on a recent project that I conducted together with Tzlil Sharon, which aimed to identify references to conspiracy theories in online text, such veiled references appear to be surprisingly pervasive (Baden & Sharon, 2021). Chiefly, there appear to be four main variants: First, allusions point at intended meanings without specifying them (e.g., "They sure got paid many Shekels for this," suggesting some anti-Semitic conspiracy theory; "I have a rope and a cozy spot…" an oblique reference to lynching), leaving it to the reader to complete the interpretation (Obeng, 1997; Wilson & Sperber, 2012). Second, language use often reaches beyond the present text into co-present contributions, using anaphora (e.g., "this," "she") to import additional meaning (e.g., "They're going to kill all the pigs in the region [to prevent the swine flu from spreading]" – "That is bad news for [German Chancellor] Merkel!"; see also Halliday & Hasan, 1976). Very similarly, multimodal communication reaches beyond the present text into co-present visual information to create additional meanings (Ben-David & Matamoros-Fernández, 2016). Third, intertextuality does the same, but reaches out to absent, supposedly familiar texts (e.g., "Die Fahnen hoch…" quoting the first words of the Horst Wessel song, anthem of the National Socialist German Workers' Party; see also Kristeva, 1981). Fourth, speakers can avoid specifying offensive meanings, using exophoric references to events, actors, or other objects in the world that are presumed to be known to other readers (e.g., posting on the day of the Christchurch terror attack: "What a great day, hopefully also soon here;" "Time to go ER," ER being the initials of Elliot Rodger, the early leader of the Incel movement and perpetrator of the 2014 Isla Vista attacks; Jaki et al., 2019).

While it is theoretically possible to algorithmically model those disambiguations needed to handle such oblique forms of hate speech (Kumaresan & Vidanage, 2019), in practice, such an enterprise quickly approaches its limits. For instance, adjacent interactive speech content can be included in the training data—but it is often unclear what nearby information an anaphora refers to (e.g., "That's way too nice" might refer to the preceding comment, the hierarchically superior comment, or the original post, each time inviting different readings; Halliday & Hasan, 1976). Likewise, it would be invalid to assume that any included anaphora refers to adjacent texts, as the same words can be used also exophorically to refer to salient present situations outside the text. Many allusions, and most intertextual references might be disambiguated by contextualizing present posts against relevant reference corpora, such as natural discourse samples on related matters, encyclopedic knowledge, or the day's news (Baden, 2018). Alas, knowing just what relevant reference corpus might be required more often than not requires that one is already familiar with a wide variety of related language uses, contextual knowledge, and recent news. Moreover, even if relevant reference corpora can be identified in an inductive fashion (e.g., by online search), machine classification still requires relevant reference materials to be labeled (Warner & Hirschberg, 2012). Clearly, continually annotating and adding any potentially relevant text to an ever-growing reference corpus, just in case that any of it might be needed to disambiguate potential hate speech, is not a viable strategy.

Even if it were possible to enable classification by considering such encompassing reference corpora, any expansion of context data shifts the detection of hateful content further away from binary, rule-based decisions toward probabilistic judgments, where both 1 (*certainly objectionable*) and 0 (*certainly harmless*) are rare occurrences. The larger the reference data, the more likely will instances be matched by pure coincidence, inflating false positive ratios (Kumaresan & Vindage, 2019). The same is true for every expansion of the textual context considered toward classification. In addition, increased reliance on reference corpora shifts the responsibility for detection away from software-controlled rulesets toward a reliance on third party algorithms (e.g., google) and patterns that emerge inductively from the reference data. Given the sensitivity of falsely redacting legitimate contents, and the consequent need for rather high classification thresholds, context-augmented machine classification is likely to achieve

at most modest improvements in detection, at considerable cost in terms of algorithmic complexity, data and labor demands, and justification.

6 The words that weren't enough

Recognizing these limitations, much current practice in content moderation continues to rely on human judgment—often following a flagging procedure that relies on any of the algorithms sketched above (Kalsnes & Ihlebæk, 2021). While much less systematic in their appraisal of available information and context, human judges should generally outperform algorithms in their capacity to detect veiled offensive content—simply because such oblique expressions are designed to be understood by humans, and missed by computers. Picking up on suspicious word choices and omissions, human judges can disambiguate allusions, intertextuality and references to adjacent texts or present situations by comprehending the context wherein a user comment was made.

And yet, even humans are often unable to decide the status of a comment—not because they cannot extrapolate those meanings expressed by the text, but because the same text supports more than one possible meaning (Boxman-Shabtai & Shifman, 2014; Warner & Hirschberg, 2012). Beyond the use of language to convey unambiguous meaning in oblique ways, the same strategies also permit the construction of properly ambiguous messages. For instance, does the comment "Someday my friends and I will come visit" convey a friend's announcement, a fan's admiration, or a veiled threat? Unless we know more about the relation between the commenter and the addressee, the statement defies disambiguation. "Why don't you go home, leave us in peace!" is ambiguous (personal/collective "you," which may/may not be a racial reference, home as home/home country, us as particular group/nationalist reference, etc.) even if we know their relation not to be close. A particularly important genre of ambiguity concerns apparent irony or humor, wherein it remains unclear whether denoted meanings are endorsed or rejected (e.g., Boxman-Shabtai & Shifman, 2014; Hodge & Hallgrimsdottir, 2020).

Using ambiguity, authors can express even meanings that are heavily sanctioned—e.g., violent threats, calls to violence, and other criminal offenses—while maintaining plausible deniability and (likely) avoiding algorithmic redaction.

Contrary to intuition,[2] such ambiguity is actually quite common in contentious discourse. For instance, in our study of conspiracist discourse, less than a tenth of all references to conspiracy theories were entirely unambiguous (Baden & Sharon, 2021). Some statements cued conspiracy theories, but left a backdoor open for benign readings (e.g., "[US Senator] Bernie [Sanders] is controlled opposition"). Others were fully ambiguous: "Nobody sued the media for creating an atmosphere like this." Of course, conspiracist discourse is known for its evasive style, as proponents of conspiracy theories have long faced social sanctions; however, the same should be true for hate speech.

One drawback of ambiguity is, of course, that the speaker's intentions may be misunderstood—a problem solved in conspiracist discourse by primarily addressing fellow believers whose predilection for certain interpretations can be safely predicted. The same logic enables ambiguous hate speech to the extent that it is intended primarily to be understood by fellow haters (Magu et al., 2017). However, to ensure that also addressed outsiders catch the intended drift, authors need to either decrease ambiguity (increasing the risk of redaction and other sanctions), or demonstratively emphasize the ambiguity, so as to alert readers to the availability of additional, hostile meanings (e.g., by adding "..." or "😉"). Unable to conclude confidently that available benign meanings were intended, the addressee is thus forced to construct and consider also the offensive interpretation.

Inversely, many cases of ambiguous statements are arguably harmless and arise accidentally when people choose their words carelessly and fail to exclude alternative, hostile meanings (e.g., when US Senate minority leader Schumer said that two Trump appointees to the Supreme Court would "pay the price" for a vote against abortion rights).[3] Consequently, flagging any speech that potentially supports offensive meanings inevitably captures numerous harmless or unintended instances, while excusing any that support harmless meanings likely misses some of the most hostile, but deliberately cloaked attacks. Especially for statements

2 When confronted with ambiguous statements, readers typically decide intuitively on one preferred reading and ignore other available interpretations, raising the illusion that most language is unambiguous. However, when prompted to make no assumptions but systematically evaluate those meanings enabled by a statement, many more statements turn out to be ambiguous (Eco, 1979)

3 https://www.washingtonpost.com/nation/2020/03/05/schumer-trump-supreme-court/

which support multiple equally plausible interpretations, there cannot be a consistent policy even for human judgment, as coders are forced to choose between redacting content that can plausibly be defended as harmless, or permitting content that can reasonably be understood as hate speech.

7 The words that were read

One final approach, accordingly, that has been widely adopted for the moderation of digital content, relies on audiences' subjective interpretations to flag offensive content. Contents get redacted, or submitted for review, if a certain number or proportion of readers regards them as offensive and flags them as such (Kalsnes & Ihlebæk, 2021). In this way, moderators can exploit the vastly superior capacity of diverse audiences to recognize oblique meanings—although at the cost of inevitably moderating post hoc, with considerable delay. However, also this strategy comes with important limitations.

To begin, especially where hateful comments are apparent only to members of extremist communities, most readers are likely to miss offensive meanings, while those who "get" the expressed hostility are likely to agree and thus unlikely to report the statement (Jaki et al., 2019). The more hate speech relies on context-based disambiguation and ambiguity, the more its detection depends on individuals who are literate in extremist discourses but in disagreement with their underlying values (see Becker & Troschke in this volume).

Furthermore, user complaint-based moderation is always vulnerable to targeted campaigning, as has been recently made salient by the rise of "cancel culture," predominantly in US-based communication forums (Ng, 2020). Given the fundamental ambiguity of language as well as the wealth of available contexts, it is very often possible to construct a statement as offensive—even if it was neither so intended nor widely understood as such. Activist users can thus use the flagging option to strategically suppress unwelcome voices wherever these miss possible ambiguities in their statements—a threat that is particular salient in the context of satire, which frequently relies on ambiguous language to confront contentious issues.

8 Conclusion

Over the course of the past two to three decades, there have been several important advances in our capacity to algorithmically detect and redact hate speech. At the same time, every advance has also revealed new limitations and contingencies in the classification of potentially offensive meanings, and provoked further adaptations in the use of evasive language suitable to express hostility in ways that are unlikely to be detected.

As I have attempted to show in this chapter, many important limitations in our capacity to detect hate speech do not primarily reflect inadequacies in those tools and algorithms employed to classify natural language, but derive from the evasive use and ambiguity of language itself (Bavelas et al., 1990). While available algorithms are increasingly capable of resolving ambiguities that exist *within* the classified text (e.g., misspellings, polysemy or different pragmatic uses; Schmidt & Wiegand, 2017), most of the remaining ambiguities reach *beyond* the text itself into intertextual context, the identities of involved actors, and the embedding social situation and communication culture (Wilson & Sperber, 2012). Of course, it is in principle possible to include ever wider context data, consider metadata information, or utilize reader reactions and talkbacks to augment classification (Baden, 2018); alas, given the vast range of potentially relevant contextual information (e.g., concurrent news, subcultural discourses, historical reference material, popular culture), including sensitive personal data (e.g., if accurate classification requires the knowledge that an addressee is gay, female, or from New York; Kumaresan & Vidanage, 2019), such an endeavor appears neither particularly practicable nor ethically defensible. Additional issues arise where detection relies on users' subjective judgments and third party-controlled data sets, and where binary decisions to permit or censor content are based on probabilistic, error-prone classifications (see Laaksonen in this volume, for a further discussion of these issues). Even if all these issues could be solved, further disambiguation is unlikely to push back the frontier by much: As in any arms race, hostile users of digital communication technologies are likely to respond to such advances by retreating deeper into the realm of ambiguous language, for which there logically cannot be an algorithmic disambiguation.

Moreover, any attempt to classify and sanction ambiguous speech is bound to raise intense contestation and public backlash (Shen & Rosé, 2019). Beyond the

inevitable rise in misclassifications, redacting comments that can be plausibly construed as harmless not only invites the justifiable indignation of the sanctioned authors, but it also sets a precedent for preemptively suppressing potentially offensive content. Accused of either censoring free speech or permitting contents that can be interpreted as offensive, neither ambiguous language, nor the black-boxed probabilistic classification can offer much grounds for justification, and even human judgment remains subjective and contestable. In light of the considerable demand on data and algorithms, the limited scope of likely improvements in detection, and the substantial damage for democratic public debates that may arise from an ill-justified suppression of ambiguous statements, attempting to pursue hate speech into the realms of evasive and ambiguous language may well do more harm than good. In terms of the arms race metaphor introduced above, an effective defense against heavily context-sensitive, evasive forms hate speech most likely requires unjustifiable infringements upon people's privacy and freedom—and where hateful communication is clad in fully ambiguous uses of language, there can be no effective defense.

Christian Baden is Associate Professor at the Department of Communication and Journalism at the Hebrew University of Jerusalem, Israel. https://orcid.org/0000-0002-3771-3413

References

Baden, C. (2018). Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In P. D'Angelo (Ed.), *Doing news framing analysis II: Empirical and theoretical perspectives* (pp. 3–26). Routledge.

Baden, C., & Sharon, T. (2021). Blinded by the lies? Toward an integrated definition of conspiracy theories. *Communication Theory, 31*(1), 82–106. https://doi.org/10.1093/ct/qtaa023

Bavelas, J. B., Black, A., Chovil, N., & Mullet, J. (1990). *Equivocal communication.* Sage.

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication, 10*, 1167–1193.

Boxman-Shabtai, L., & Shifman, L. (2014). Evasive targets: Deciphering polysemy in mediated humor. *Journal of Communication, 64*(5), 977–998. https://doi.org/10.1111/jcom.12116

Eco, U. (1979). *The role of the reader: Explorations in the semiotics of texts.* Indiana University Press.

Fairman, C. M. (2006). Fuck. *Public Law and Legal Theory Working Paper Series, 59*. Ohio State University.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society, 20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gruber, H. (1993). Political language and textual vagueness. *Pragmatics, 3*(1), 1–28. https://doi.org/10.1075/prag.3.1.01gru

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Routledge.

Hodge, E., & Hallgrimsdottir, H. (2020). Networks of hate: The alt-right, "troll culture", and the cultural geography of social movement spaces online. *Journal of Borderland Studies, 35*(4), 563–580. https://doi.org/10.1080/08865655.2019.1571935

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict, 7*(2), 240–268. https://doi.org/10.1075/jlac.00026.jak

Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society, 43*(2), 326–342. https://doi.org/10.1177/0163443720957562

Kristeva, J. (1981). *Language and desire: A semiotic approach to literature and art.* Columbia University Press.

Kumaresan, K., & Vidanage, K. (2019). HateSense: Tackling ambiguity in hate speech detection. *Proceedings of the IEEE 2019 National Information Technology Conference,* 20-26. https://doi.org/10.1109/NITC48475.2019.9114528

Magu, R., Joshi, K., & Luo, J. (2017). Detecting the hate code on social media. *Proceedings of the International AAAI Conference on Web and Social Media, 11*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14921

Mufwene, S. S. (2001). *The ecology of language evolution.* Cambridge University Press.

Ng, E. (2020). No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media, 21*(6), 621–627. https://doi.org/10.1177/1527476420918828

Obeng, S. G. (1997). Language and politics: Indirectness in political discourse. *Discourse & Society, 8*(1), 49–83. https://doi.org/10.1177/0957926597008001004

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Searle, J. R. (1969). *Speech acts.* Cambridge University Press

Shen, Q., & Rosé, C. P. (2019). The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. *Proceedings of the Third Workshop on Abusive Language Online*, 58–69. https://doi.org/10.18653/v1/W19-3507

Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, *22*(12), 2218–2237. https://doi.org/10.1177/1461444819888746

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the 2012 Workshop on Language in Social Media,* 19–26.

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Zelenkauskaite, A., Toivanen, P., Huhtamäki, J., & Valaskivi, K. (2021). Shades of hatred online: 4chan memetic duplicate circulation surge during hybrid media events. *First Monday, 26*(1). https://doi.org/10.5210/fm.v26i1.11075