

“Somewhere along your pedigree a bitch got over the wall!” – A data-driven approach to a typology of implicitly offensive language

Kristina Š. Despot^a, Ana Ostroški Anić^a, Tony Veale^b

^aInstitute for the Croatian Language

Republike Austrije 16, 10000 Zagreb, Croatia

^bUniversity College Dublin

Belfield, Dublin 4, Ireland

Kristina Štrkalj Despot (**Corresponding author**)

kdespot@ihjj.hr

<https://orcid.org/0000-0001-9004-5103>

Ana Ostroški Anić

aostrosk@ihjj.hr

<https://orcid.org/0000-0001-9999-0750>

Tony Veale

tony.veale@ucd.ie

<https://orcid.org/0000-0003-2375-1811>

“Somewhere along your pedigree, a bitch got over the wall!” –

A data-driven approach to a typology of implicitly offensive language

ABSTRACT

The automatic detection of implicitly offensive language is a challenge for NLP, as such language is subtle, contextual, and plausibly deniable, but it is becoming increasingly important with the wider use of large language models to generate human-quality texts. This study argues that current difficulties in detecting implicit offense are exacerbated by multiple factors: (a) inadequate definitions of implicit and explicit offense; (b) an insufficient typology of implicit abuse; and (c) a dearth of detailed analysis of implicitly offensive linguistic data. In this study, based on a qualitative analysis of an implicitly offensive dataset, a new typology of implicitly offensive language is presented along with a detailed, example-led account of the new typology, an operational definition of implicitly offensive language, and a thorough analysis of the role of impoliteness triggers, figurative language, and humor in each type. Our analyses identify three main issues with previous datasets and typologies: (a) conflating content and form in the annotation; (b) treating figurativeness, particularly metaphor, as the main device of implicitness, while ignoring its equally important role in the explicit offense; and (c) an over-focus on form-specific datasets (e.g. focusing only on offensive comparisons), which fails to reflect the full complexity of offensive language use.

Keywords: implicit offense, automatic detection of implicitly offensive language, large language models, metaphor, humor, impoliteness triggers

Acknowledgements

We are thankful to Luka Terčon for his help with the definition and decision tree on the implicit vs. explicit offense, and to Barbara Lewandowska-Tomaszczyk, the leader of the Use Case 4.1.1. *Incivility in Media and Social Media* within the CA Nexus Linguarum for motivation and support.

Funding sources

This work was supported by the COST Action COST Action CA 18209 - *European network for Web-centered linguistic data science Nexus Linguarum*

“Somewhere along your pedigree, a bitch got over the wall!” –

A data-driven approach to a typology of implicitly offensive language

ABSTRACT

The automatic detection of explicitly offensive language is a reliable feature of many NLP (natural language processing) systems. However, the detection of implicitly offensive language still remains a challenge, as such language is more subtle, more contextual, and more plausibly deniable. Indeed, because creative taboo-breaking and implicit offense is a widespread feature of human language, the detection of implicitly offensive language is becoming increasingly important with the wider use of large language models (LLMs) to generate human-quality texts. This study argues that current difficulties in detecting implicit offense are exacerbated by multiple factors: (a) inadequate definitions of implicit and explicit offense; (b) an insufficient typology of implicit abuse; and (c) a dearth of detailed analysis of implicitly offensive linguistic data. Existing typologies continue to be used due to their utility in earlier work, or because of their perceived overlap with existing datasets (e. g. van Aken et al. 2018, Wiegand et al. 2021, ElSherief et al. 2021). Here we use a dataset acquired for an EU COST action, the *European network for Web-centered linguistic data science*. The data comprises 331 documents that were annotated for offensiveness with the INCEpTION tool (Lewandowska-Tomaszczyk et al. 2021) by a pair of annotators and a curator. Based on a qualitative analysis of 200 sentences that were tagged as *implicitly offensive* (one of which we use in the title), a new typology of implicitly offensive language is presented. This typology differentiates between the nature of the offense (aggressive speech, insulting speech, discrediting/condescending speech, dehumanization, derogation, and discrimination), and the linguistic devices used to express it (metaphor, metonymy, simile, irony, hyperbole, euphemism, repetition, rhetorical questions, circumlocution, name calling, contrastive statements, and the use of graphic devices). In addition to a detailed, example-led account of the new typology, we provide an operational definition of implicitly offensive language and a thorough analysis of the role of impoliteness triggers, figurative language, and humor in each type. Our analyses identify three main issues with previous datasets and typologies: (a) conflating content and form in the annotation; (b) treating figurativeness, particularly metaphor, as the main device of implicitness, while ignoring the equally important role of metaphor in explicit offense; and (c) an over-focus on formal criteria, or on form-specific datasets (e.g. focusing only on offensive comparisons), which fails to reflect the full complexity of real offensive language use.

Keywords: implicit offense, automatic detection of implicitly offensive language, large language models, metaphor, humor, impoliteness triggers

1. Introduction

In our digitally-connected societies, we now have greater access to communication platforms that offer faster and more pervasive modes of linguistic expression. This increased freedom of expression allows marginalized groups to voice their concerns and opinions, but it has also fostered an increase in public displays of aggressive and harmful language, such as through verbal aggression, taboo-breaking, hate speech, bullying, harassment, and badgering (Patrick and Hollis 2018). These egregious forms of speech can be seen as a violation of social norms of communication, particularly of the Gricean cooperative principle (Grice 1975). Although these behaviors are scarcely new to modern culture, and have always existed in various guises throughout history, the widespread availability of social media technologies expands their reach and magnifies their effects, which has led to increased efforts to mitigate their worst outcomes. Offensive language, whether explicit or implicit – through euphemism, metaphor, or “dog whistles” – can cause emotional distress, incite hatred, damage relationships, and escalate conflicts to the point of physical violence.

Offense takes root in many forms, and so computational approaches to its automated detection speak of different forms in different terms, whether *incivility* (Sadeque et al. 2019, Stoll et al. 2020), *toxic language* (Kunupudi et al. 2020), *abusive language* (Caselli et al. 2020, Waseem et al. 2017, Wiegand et al. 2021), *offensive language* (Zampieri et al. 2019) or *hate speech* (Gao et al. 2017, ElSherief et al. 2021, Schmidt and Wiegand 2017). Nonetheless, there is general agreement that offense through language has both explicit and implicit forms, and that its target may be quantifiable or amorphous, as either a specific individual or a group. Lexical (and lexicon-based) approaches to detecting explicitly offensive language have made headway by focusing on the specific words, while the detection of implicit offense has proven far more challenging, not least because this kind of offense is designed to toy with our own internal censors, nevermind those of machines. Explicit offense wears its intentions openly; implicit offense requires a typology of forms and methods to better understand its hidden intentions.

Moreover, the detection of implicitly offensive language is growing in importance with the widespread adoption of large language models (Li, 2022), as e.g. used in the chatbot *ChatGPT* (Leiter et al. 2023). A language model (LM) is, at its core, a probability distribution over the strings of a language; that is, a mechanism for assigning probabilities to different sequences. From a finite number of observations and with a finite number of parameters, they derive their ability to say that one string is more likely than another. The piecewise fashion in which they finitely model the infinite makes them excellent predictors of language: they can fill in blanks, or predict the most probable continuation to strings that they (and us) have never seen before.

When the occasion calls for explicit offense, LMs are quick to oblige. For instance, a LM will assign a high probability to the expletive that most obviously fills the gap in this masked utterance: “I thought I told you to ___ off!” LMs do not have different modules for syntax, semantics and pragmatics, but conflate different factors in language production into a single end-to-end mapping of inputs (text prompts) to outputs (text continuations). LMs learn to play language games from the large datasets on which their parameters are tuned. This makes them savants at the language of smalltalk, formal speeches, literary genres, computer coding, and math problems. Their ability to complete our thoughts by completing our sentences means that they can also cleverly veil offensiveness with implicitness, just by learning the tacit rules of the language games that we humans play whenever *we* wield words like hidden daggers.

Large LMs, the so-called LLMs, have many more parameters, are trained on ever larger datasets, and use wider context windows for maintaining coherence across a discourse. They are also capable of greater abstraction and more powerful generalizations. We might expect an LLM to learn how to “talk the talk” of solving puzzles or analyzing poetry, but they seem able to “walk the walk” in these domains too. For instance, ChatGPT can offer a detailed analysis of novel metaphors such as “How is Tom Hanks the potato of the acting world?” (answer: Hanks is a popular, versatile and “down-to-earth” actor). When asked to solve the metaphor riddle “So what is the Donald Trump of the fruit world?”, the LLM does not opt for the superficial answer (an orange), but suggests instead the durian fruit, pointing to this fruit’s “polarizing reputation, with some people loving its unique taste and others finding its odor and flavor overpowering.” Interestingly, ChatGPT at first demurs in its response, stating that the given prompt is not a commonly-used phrase or metaphor in its training data. However, when pushed to treat it as a novel comparison, it offers an analysis that is concise and incisive.

This demurrer arises in part from the model’s desire to avoid the appearance of bias, which it achieves with mixed success through the implementation of input “guard rails.” There are credible fears that LLMs may be weaponized by bad actors to generate hateful content on an industrial scale. This makes the automated detection of offense, both explicit and implicit, a prime concern for public-facing LLMs such as ChatGPT. Consider the model’s analysis of the phrase in this paper’s title. It correctly interprets the word “bitch” to mean “an individual who was not from the same social class or racial background [who] managed to enter the family bloodline, perhaps through a romantic relationship or an affair”, and goes on to note that the word “is particularly derogatory, as it implies that the individual [...] was a woman who was seen as being of lower social status or moral character.” It concludes that “making derogatory

comments about someone's ancestry or heritage is not acceptable behavior.” Those who *talk the talk* of sensitivity and respect do not always *walk the walk*, but ChatGPT shows us that a productive facility for the former is a useful platform on which to build the latter.

This paper explores the limitations of current methods and of existing typologies for detecting implicit offense in language. On the basis of our study, we argue that difficulties chiefly stem from (a) an inadequate definition of implicit vs. explicit offense; (b) an imprecise typology of implicit offense; and (c) a lack of a detailed linguistic analysis of implicitly offensive datasets. We suggest several ways of overcoming these difficulties. We first survey how related work has defined explicit and implicit offensive language, and examine existing typologies for classifying the latter. We then suggest an operational definition of implicit offense and present a new typology to organize its various types. Each type of offense in the typology is illustrated with examples from our dataset, the analysis of which will highlight the role of impoliteness triggers, figurative language, and humor when conveying offense. Finally, we draw attention to aspects that prove challenging for machine-learning approaches to detection, and suggest ways in which these challenges can be mitigated with finer-grained typologies and datasets. An additional question that has motivated us and that we tried to answer is whether modern NLP technologies (e. g. public-facing LLMs such as ChatGPT) can detect and grasp the full richness of the ways these devices are exploited in our data.

2. Related work

2.1. An overview of the definitions of implicitly offensive language in NLP research

Waseem et al. (2017) were among the first to call for more work to be done on the detection of implicit abuse in language, which they define as “that which does not immediately imply or denote abuse” (Waseem et al. 2017: 79). In contrast, explicit language is “unambiguous in its potential to be abusive” (ibid), as is evident from e.g. racial or homophobic slurs. Kumar et al. (2018) prefer the terms *overt* and *covert aggression*. Their approach is one of many that makes a formal distinction between both types of abuse by reference to a lexicon of offensive words. If a text contains “more directly offensive words than potentially offensive words” (Kumar et al. 2018: 7), it is marked as overtly aggressive, while texts containing some potentially offensive words are marked as covert aggression. Gao et al. (2017) identify circumlocution, metaphor, and stereotypes as specific linguistic devices for conveying ill will through implicit hate speech,

“in which hatefulness can be captured by understanding its overall compositional meanings” (2017: 775).

Caselli et al. (2020) aim for a subtler differentiation between explicit and implicit abuse, and define implicit abuse as that which can only be “suggested or inferred” because “it does not have any surface evidence” (2020: 6197). For these authors, explicit abuse is manifested by means of “profanity, performative constructions, imperatives, idioms, adjectives, or nouns with a clear negative connotation.” In contrast, the devices used to veil abuse in implicitness include “sarcasm, metonymy, irony, litotes, euphemism, and inside jokes” (ibid). Although *negative connotation* is not otherwise defined, we infer from the given examples that it relates to words and expressions that convey a critical opinion or attitude about a target.

ElSherief et al. (2021) define implicit hate speech as “a subclass of hate speech defined by the use of coded or indirect language such as sarcasm, metaphor, and circumlocution.” Similarly, in their typology of implicit abusive language, Wiegand et al. (2021) also aim for a practical definition by understanding implicit abusive language as offense that is “not conveyed by (unambiguously) abusive words (e.g. dumbass, bimbo, scum)” (2021: 576). By lexicalizing explicitness, they consign critical statements whose abuse is only apparent at the phrase level to the class of implicit abuse, such as *I haven't had an intelligent conversation with a woman in my whole life.* (Wiegand et al. 2021: 576), *Women should all stay at home in an apron, chained to the oven!* and *Get back in the kitchen.* (Wiegand et al. 2021: 578).

From a computational perspective, explicit abuse is therefore the kind of offense that easily can be captured with a “blacklist” of undesirable terms. (We note that the term “blacklist” is now itself a proscribed term, with “blocklist” becoming the preferred alternative.) But the fragility of this perspective has long been understood by comedians. In his routine “the seven words you can't say on TV,” George Carlin famously highlighted the existence of “two-way words” that have both acceptable and unacceptable uses. As he put it in his characteristically salty manner, “You can prick your finger ... Just don't finger your prick” (see Author 3, 2021).

2.2. Existing typologies of implicitly offensive language in NLP

Two recent typologies of implicit offense are relevant to any effort to define the phenomenon. The first is by Wiegand et al. (2021), who analyzed eight English datasets as part of a binary classification of microposts as abusive or otherwise. These authors calculated the proportion of explicit abuse by detecting the presence of abusive words from the lexicon of Wiegand et al.

(2018). After manually annotating a random sample of 500 implicitly abusive microposts from each dataset, they proposed eight subtypes to characterize the implicit abuse in the data: 1) stereotypes; 2) perpetrators, a “morally-contemptible” sibling of stereotypes to which egregious and often criminal behaviors are commonly (if unfairly) attributed (Wiegand et al. 2021: 578); 3) abusive comparisons; 4) dehumanization, or the “act of perceiving or treating people as less than human” (2021: 579) by means of unfavorable comparisons with animals and other non-human tropes; 5) euphemistic constructions, often expressed using negation, such as “I wouldn’t touch [target] with a ten-foot pole”; 6) calls for action, where an author appeals for violence against the target); 7) multimodal abuse, i.e. abusive content presented via non-textual elements (e.g., visual memes, emoji); and 8) phenomena requiring world knowledge and inferences, which uses inference over shared knowledge in jokes, sarcasm, rhetorical questions, topical allusions and so on to communicate a criticism that is implied by a speaker and inferred by the reader.

When building their Twitter dataset of implicit hate speech, ElSherief et al. (2021) applied both fine-grained labels and, significantly, free-text descriptions of the implied aspects in each post. Their taxonomy of labels defines six classes of implicit hate speech that are “grounded in the social science literature” (ElSherief et al. 2021: 346). They define the class *White Grievance*, as evidenced in the tweet *Black lives matter and white lives don’t? Sounds racist*, as “frustration over a minority group’s perceived privilege and casting majority groups as the real victims of racism” (2021: 347). The class *Incitement to Violence* includes displays of in-group unity and out-group rejection in the service of hateful ideologies, such as Nazism. The class *Inferiority Language* relates to acts of dehumanization and toxic behavior in the furtherance of power disparities, while the class *Irony* unites sarcasm, satire, and any form of humor that is used to “attack or demean a protected class or individual” (2021: 347). The class *Stereotypes and Misinformation* labels instances of hate speech that exploit and reinforce the negative qualities unfairly ascribed to a protected class, while *Threatening and Intimidation* denotes the speaker’s willingness to embrace violence or other, subtler forms of intimidation.

These classes are used to label tweets in a second pass of annotation, after an initial pass has determined whether a given text is an example of explicit hate, implicit hate, or no hate at all. The methodology, which uses crowd-sourced annotations for this two-stage process, can be replicated on other datasets, and its work products can be used in the supervised training of machine learning models. The six classes, which are not mutually exclusive, are chosen to

model the hate ideologies that are most prevalent in the US, and must likely be expanded to cover the kinds of implicit speech one might encounter in other cultural contexts and datasets.

3. A data-driven approach to implicit offensive language typology

These typologies emerge from a mix of top-down and bottom-up approaches, in which classes of implicit offense are postulated on the basis of their utility in previous work, or on their observed frequency in the datasets being analyzed (e. g. van Aken et al. 2018, Wiegand et al. 2021, ElSherief et al. 2021). The classes are typically considered self-evident, and detailed origin stories for each class are often omitted. Moreover, even though some classes are shared by different approaches, the typologies of these approaches can still differ considerably.

This brief survey shows that researchers can also differ widely in their working definitions of implicit offense. Since implicitness is such an elusive category, a solid definition is a crucial first step, as underspecified or unclear definitions lead to confused and unreliable annotations. Earlier definitions fall broadly into two groups: those driven by emotional-valence (is the negative valence unambiguous and explicit, or deniably implicit?), and vocabulary-driven (the overt presence of abusive words makes an offensive intent explicit, while an offensive intent that avoids any such words remains implicit, e.g. see Wiegand et. al. 2021).

The valence of a text is not always self-evident, and it can be hard to achieve a satisfying level of agreement between different annotators. We thus rely on a vocabulary-driven definition that is more amenable to manual annotation and to machine-learning tasks. In this view, an explicitly offensive example is one that contains words designated as offensive in a dictionary, and that are undoubtedly aimed at the target. Conversely, if a text does not use a single lexical item that unambiguously denotes hatred or malicious intent towards a target, then any offense must be implicit. A basic threshold must be passed, then: a text must have at least one lexical item that, according to the relevant dictionary, has at least one sense marked as offensive, insulting, vulgar, or similar, and this exact sense must be used in the example. Implicitness in this sense is not to be equated with indirectness (Searle 1979), but as will be evident from the analysis of various examples, they are tightly connected ideas since, in many cases, implicitly offensive language makes use of indirect speech acts, presuppositions and implicatures.

We begin by annotating texts that have already been pre-labeled as implicitly offensive. We then classify these short texts into different types of offensiveness via an in-depth qualitative

analysis of each one. This analysis takes account of the communicative strategies and triggers of impoliteness (i.e. conventionalized impoliteness formulae) that can cause social conflict and disharmony. To this end, we make reference not just to impoliteness theory (Culpeper 2011, Culpeper 2016), but to the mode of figurative language analysis used in Dancygier and Sweetser (2014) and Author 1 et al. (2019), and to various theories of humor surveyed in Author 3 (2021). The resulting typology emerges bottom-up from the types of implicit offense found in the examples when they are viewed through these various lenses.

Devices such as metaphor and simile, and even irony, can be used with or without offensive intent. Indeed, even aggressive speech can be used for self-mockery or harmless parody. It is thus crucial that we do not conflate the device used in a text with its ultimate intent, and so we propose a multi-level typology that differentiates between the goal of an offense (aggressive speech, insulting speech, discrediting/condescending speech, dehumanization, derogation, and the discriminatory use of stereotypes, racism, sexism, etc.) and the linguistic devices used to achieve it (metaphor, metonymy, simile, irony, hyperbole, euphemism, repetition, rhetorical questions, circumlocution, name-calling, explicit contrasts, graphic devices, etc.). Metaphor, for instance, can be used to elevate or to diminish¹. We use it to personify the non-human and to dehumanize our fellow humans. Its myriad uses may be subtle or crass, implicit or explicit. It would be inappropriate, then, to consider dehumanization and metaphor as sibling classes in any taxonomy. Rather, metaphor is just one means by which dehumanization is achieved, and we must be careful to separate the means from the ends in our proposed typology of offense.

3.1. The dataset and methodology

For this study, we have used data annotated within the WG 4 of the COST Action CA 18209 - *European network for Web-centered linguistic data science Nexus Linguarum* (Use Case 4.1.1. *Incivility in Media and Social Media*). Within this use case, 30 previously annotated offensive language datasets were used to annotate randomly extracted documents by applying a more detailed typology for offensive language annotation (Lewandowska-Tomaszczyk et al. 2021). A total of 331 documents, containing 1264 sentences, have been annotated at multiple levels by two annotators and one curator. The first level marks whether the intent of an offensive utterance appears to be expressed explicitly or implicitly. We then collected 200 sentences that

¹ For an overview of metaphor research in the field of cognitive linguistics see Dancygier and Sweetser 2014.

had been marked as implicitly offensive in this way by at least one annotator. Those sentences were then independently annotated by two other annotators for the nature of the offense and for the linguistic device that was used to convey it.² The broader context, including the whole document if necessary, may be taken into account when adding these later annotations. Both annotators then compared their results, discussed points of disagreement, and found common labels for examples that were similar in content. The completed dataset of annotated sentences is available from <https://tinyurl.com/ymaff9pu> to view and download.

3.2. Types of implicitly offensive linguistic units derived from the data

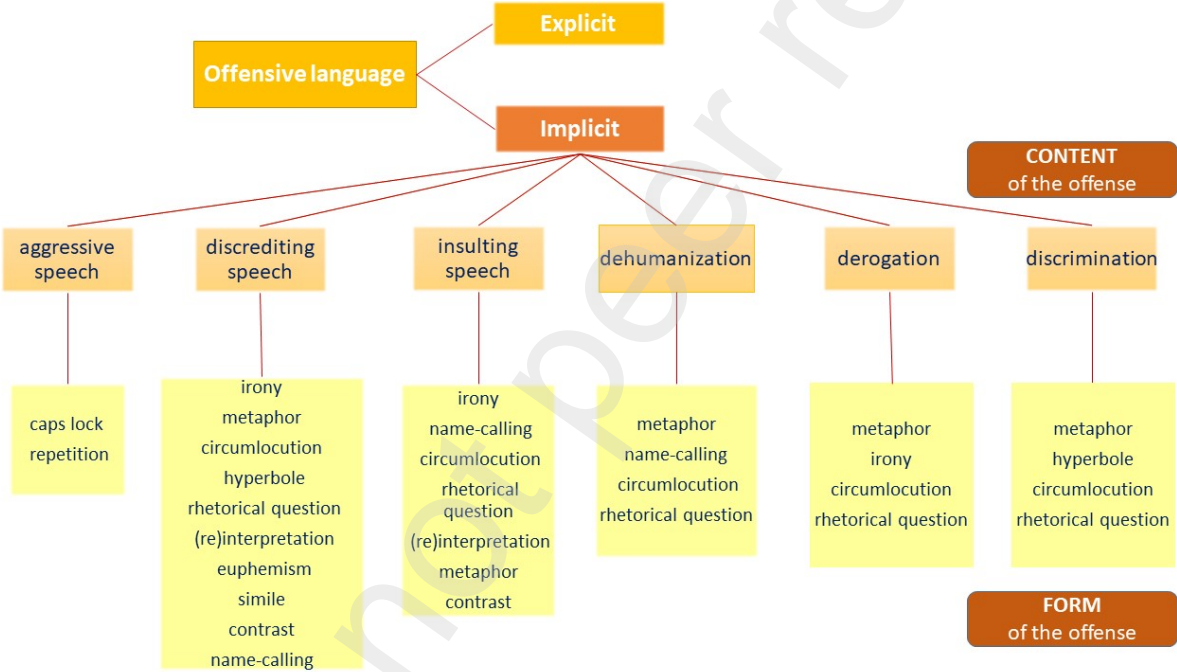


Figure 1. *Types of implicitly offensive linguistic units derived from the data*

3.2.1. Aggressive speech

In its implicitly offensive form, aggressive speech is the textual equivalent of yelling, hissing and snarling. It is marked by its confrontational tone and its intent to intimidate. This type of language is often used to assert dominance or to express anger, frustration, or dissatisfaction.

² Both annotators are experienced in metaphor annotation tasks (see Author 1 et al. 2019).

In our dataset, aggressive speech is realized through linguistic features such as repetition (example 1 below), all-caps spelling (examples 3 and 4 below), and ill-tempered rhetorical questions that often carry veiled threats (example 1, 3 and 4). A sense of intensity and urgency is created using punctuation (example 2) and repetition (example 1) – particularly of vowels (example 2), in ways that can be interpreted as impatient and impolite. Conversely, the use of all-caps is the textual equivalent of shouting to loudly and aggressively emphasize a point.

- (1) Is humiliation or hurt a likely response to that, Malleus? Huh? Huh?
- (2) Noooo!!! What a shock!
- (3) Should I dismiss EVERYTHING because Hitler bad?
- (4) Does any OTHER admin want to come and delete this idea for the sake of some shallow excuse like a banned user?

Example (3) aims to counter a reference to Hitler with an aggressive form of prototype-based reasoning (Lakoff 1987). In Example (4), the metaphorical expression *some shallow excuse* achieves a discrediting effect through the conceptual metaphor UNDERSTANDING IS DEPTH.

- (5) How dare you? Just as I mentioned on his page, you are just like one of those cowardly countries that did not recognize Abkhazia and South Ossetia, but ignorantly recognized Kosovo.

Example (5) begins with an implicitly threatening rhetorical question, *How dare you?*, and is followed by an impoliteness trigger identified in (Culpeper 2016). A personalized simile is used to discredit the addressee by asserting similarity to people who have acted inconsistently and, by implication, hypocritically in the past. The offensiveness is designed to be recognized even by those who lack any historical context, via the use of the negatively valenced words “*cowardly*” and “*ignorantly*”. Similes that directly target an interlocutor by using a negatively valenced source domain are always offensive (see Wiegand 2021b, Quadir et al. 2015).

3.2.2. Discrediting/condescending speech

As a form of implicit abuse, discrediting speech aimed at an interlocutor/target attempts to undermine the credibility, competence, or integrity of their actions or words by implying that they are not intelligent, reasonable, well-mannered, trustworthy, reliable, consistent and so on. The effect can be manifested in various linguistic modes, from irony and sarcasm to rhetorical questions that challenge the interlocutor's judgment and motives, or through similes and metaphors that emphasize the interlocutor's inferiority or lack of credentials. Discrediting

speech can erode trust, damage relationships, and devalue the interlocutor's contributions. It can also be used as a power play by a speaker who seeks to establish their dominance or authority over the interlocutor. Discrediting speech can also be intended to silence others, as when a speaker mockingly questions another's sincerity or seriousness (Caponetto 2021) with utterances in the vein of "surely you can't be serious" and "are you f**king kidding me?".

Discrediting speech is so commonplace it has given rise to formulaic discrediting expressions, as evidenced in examples (6) and (7) from our dataset. The fixed nature of these expressions also make them attractive (and easily learnt) from a computational point of view.

- (6) Anyway, that's off the point.
- (7) You have got to be kidding.

The fixed expression *off the point* in example (6) uses the conceptual metaphor IMPORTANCE IS CENTRALITY to discredit the interlocutor's words as unimportant or irrelevant to the discussion. The form [that][is][off the point] follows Culpeper's (2016) conventionalized impoliteness formula of pointed criticism, where the third element uses metaphor to satisfy the model's expectation of a negatively valenced claim. Example (7) is the typical case of sarcasm, where the speaker pretends that the interlocutor is joking so as to portray their words as risible. This use of discrediting speech is dubbed *seriousness silencing* by (Caponetto, 2021).

- (8) I wonder if he'd have the same position for a riot at a Linda Sarsour speaking event...
- (9) Well, if he's SO BUSY, why did he have enough time to repeatedly delete my posts?
- (10) You wouldn't know prosperity if you fell into a pile of 100's.
- (11) It's obvious (((David))) has never read one word written by poor old sweet-tempered Chuck He's a Lover Not A Fighter Darwin.
- (12) Nazi bad, Trudeau nazi therefore Trudeau bad. Is that all the thinking that's required?
- (13) Their egos are based in quicksand.

Examples (8) and (9) can be viewed as further instances of seriousness silencing insofar as they force the inference that the interlocutor is untrustworthy, inconsistent, dishonest, or all of these things. Example (10) packages an insult in the form of personalized negative assertion (Culpeper 2016). Here the desired inference is not so much forced as explicitly stated: *You do not know what real prosperity is*. Since the interlocutor is claimed to be using words that they do not understand, they should be ignored, making this another case of seriousness silencing (Caponetto 2021). Example (11) is especially complex, and draws on a number of discrediting stratagems. The speaker not only employs metonymic hyperbole (*never read one word*), but

also over-familiarity (*Chuck for Charles*), frivolous metaphors (describing Darwin as *He's a Lover Not A Fighter*), and long-winded circumlocution (*poor old sweet-tempered Chuck ... Darwin*) to imply that the interlocutor is unqualified and ill-informed on the subject. These stratagems unite to emphasize the gulf between the knowledgeable speaker and an ignorant interlocutor. In example (12), the speaker employs a *Me Tarzan You Jane* speaking style to mock the interlocutor's lack of sophistication. The interlocutor may not speak this way, but the point – metaphorically made – is that they certainly think this way. The point is hammered home by a rhetorical question that cements the speaker's dim view of their thought processes. Example (13) taps into the conceptual metaphor COMPLEX SYSTEMS ARE BUILDINGS (BASE IS THE FOUNDATION) to suggest that the interlocutor lacks the intellectual foundations on which to build such a grandiose self-image. The use of the source concept “quicksand” also implies a threat, by suggesting that the interlocutor is not smart enough to see the danger they are in.

Irony often uses the pretense of extreme politeness, intimacy or compassion to discredit another, by implying that an interlocutor needs to be treated with great delicacy, as one might treat a child or a self-regarding fool. The pretense must be exaggerated so as to be penetrated by the audience. One may be over-kind to be cruel, and one can be so polite as to be impolite (see Culpeper 2016 on the impoliteness meta-strategy of sarcasm or mock politeness).

(14) I like that notion. My dearest sir, you are so correct!

(15) He is so sweet.

(16) How very sad.

So in examples (14)-(16), the speaker's true intent is at odds with the sentiments they express.

Discrediting can also be achieved by a forced reinterpretation of other's words or actions:

(17) I thought we were here to work on an encyclopedia, but you just seem to be here to stir up as much drama as possible.

(18) Less than 100 edits to articles in 2.5 years, yet you have no problem chiming in and rudely making bad situations worse.

Examples (17) and (18) are, following from the framework of Culpeper (2016), personalized negative assertions. Each uses a similar structure to convey a clash of perspectives: the main clause expresses the speaker's view of reality as a self-evident truth, before a contrastive conjunction introduces a critical view of the interlocutor's actions, which is also offered up as

a truth. Each example also relies on a metaphorical expression (*stir up drama, chime in*) to discredit the interlocutor's motives (or credentials) for doing what so displeases the speaker.

To discredit another person and their ideas, a speaker might also offer a forced choice between a self-evidently positive worldview and a self-evidently negative worldview. In this way, the speaker makes it abundantly clear which side they, and the interlocutor, are to be found on:

- (19) We as a country need to decide if we want individual liberty more than an overbearing centralized government telling us how to live and wasting all our money on bureaucracy.

In example (19), the speaker brings the entire “country” into the debate, making it clear that the interlocutor is not just a discredit to themselves, but to the nation; in other words, they are “a national disgrace.” Moreover, the metaphorical terms *overbearing, centralized, telling us, and wasting money* only reinforce the speaker's critical view of the interlocutor's views.

In the following examples, another person is discredited as being rude and impolite:

- (20) I don't like when people are biased and rude.
(21) Such crass, rude, and low-class behavior.
(22) As said by Sir Thomas More: They'll think that somewhere along your pedigree a bitch got over the wall!
(23) Seems ghetto manners have rubbed on Harry big time.

Example (20) is offensive if, like that of (21), it is intended to target the interlocutor. In that case, the implication employs the GENERIC IS SPECIFIC schema to place that specific interlocutor in a generic class of people that are deserving of criticism. The implied offense of example (21) is perhaps simpler, even if it also relies on circumlocution, because it offers a cumulative sequence of negatively valenced words: *crass, rude, and low-class*. Example (22) is somewhat more complicated, for us as analysts and for machines as automated detectors, as it quotes a memorable line attributed to Thomas More in the film *A Man for All Seasons*. The line draws on the conceptual metaphor CLASS IS BREEDING to describe a target's parentage in animal terms. The key term is not so much “bitch” – which has a negative valence when used of humans – but “over the wall”, which implies that the interlocutor's family line has been compromised by undesirable outside influences. As in examples (21) and (22), it is left to us to infer that the line refers to the interlocutor, to create what Culpeper (2016) calls a personalized negative reference, and this ambiguity adds to the implicitness of the offense. Nonetheless, there is a sexist dimension to the slight, since it is implied that the baleful influence is female. Example (23)

smacks of both sexism and racism in its use of the expression “ghetto manners,” which metaphorically equates the target, “Harry,” with the typical denizens of the typical ghetto. If the target is Prince Harry, the racism is magnified, insofar as it implies these manners were acquired from his wife, an actress of mixed race who is a frequent target of racist jibes.

Because it is often difficult to distinguish condescending speech from discrediting speech, we view them as a single composite category. Condescending speech uses or evokes a patronizing or superior tone to imply that the interlocutor is somehow inferior or inadequate, and the effect is always to discredit its targets. From their superior position of presumed authority, the condescending speaker may offer unsolicited advice, deliver lectures, judgments or sermons, or identify specific failings in the interlocutor. This power imbalance can contribute to feelings of frustration and resentment on the part of the intended target.

- (24) This is the worst attempt at a dog whistle that I have ever seen, we all know what she really meant: the only people you should be arresting are poor, black, men.
- (25) Don't justify murder by saying it was never alive or it's my right.
- (26) In other words, you intend to reward unproductive or irresponsible parents and penalize good parents.
- (27) You confuse science with censorship.
- (28) Allow me to spell it out for you.
- (29) But he's not an Australian, so fix it up and get your facts right.

Examples (24) through (29) each lecture their targets from a position of assumed authority. Example (24) implicitly accuses its target of deceitful behavior, by suggesting that their views are more discreditable than they would care to admit. Just as dog whistles are heard only by dogs, a speaker who engages in “dog whistle” communications uses coded language to mask the worst aspects of their social or political views, which may be sexist, racist or reactionary. The dog whistle itself is a violation of the Gricean maxim of quality (Grice 1975), but here the target is further accused of incompetence, insofar as their efforts are so nakedly transparent. It sharpens the implied offense somewhat to imagine the target’s intended audience as “dogs.”

Example (25) reinterprets the target’s views as a morally unjustifiable defense of murder. That the speaker feels qualified to make this judgment, which effectively casts the target as an apologist for crime, puts them in a self-appointed position of moral superiority. To achieve a greater impact, the speaker uses two heavily valenced metaphors: “justify,” which activates the frames of *justice* and *guilt*, and the more hyperbolic “murder,” which activates the frames of

evil, crime, and immorality. The act of reinterpretation is more explicit in Example (26), since the speaker openly admits to rephrasing the interlocutor's thoughts "in other words." These other words recast the disputed ideas in a form that is likely oversimplified, so as to discredit the target for defending such a self-evidently illogical position. Example (27) also expresses the interlocutor's position in other words, but as in (26), these words are unlikely to be ones that a target would agree with. We can see that implied offense by reinterpretation operates by replacing the interlocutor's own words with much more disobliging alternatives. The condescending presumption, which is clearly articulated in Example (28) as *allow me to spell it out for you*, is that the target cannot express themselves clearly without the help of a superior thinker. The conceptual metaphor at work in Example (28), INTELLIGENCE IS LITERACY, adds to this implied offense by suggesting that the target even needs help with their spelling. By overriding a target's own words, these offenses may also be seen as attempts at silencing, which can in turn be interpreted as an impoliteness trigger in the mold of (Culpeper 2016). Example (29) is a more obvious example of discrediting speech with a condescending tone, in which the interlocutor is chastised as an errant child that should go and fix their own mistakes.

3.2.3. Insulting speech

Insulting speech shares many similarities with discrediting speech, insofar as the end-result is the diminishment of its target. However, insulting speech is also highly personal; it aims to do more than discredit an arbitrary target as a fellow disputant, but aims to undermine them as people. The speaker is specific in their insults, using personal knowledge of the target in their attack. An insult of the form "even a moron like you realizes that ..." may even acknowledge a target's correctness on a particular issue. In contrast, discrediting speech always aims to undermine a target's credibility or competence, but it is not always personal. This personal dimension can add to the offensiveness of insulting speech when it cuts closer and deeper.

Our view on insulting speech gels with Culpeper's (2016) impoliteness trigger of insult, which identifies personalized negative vocatives, personalized negative assertions, personalized negative references, and personalized third-person negative references as common devices for implicitly insulting language. The personal dimension may also be exploited by using names, especially nicknames and diminutives, since these are often used by intimate acquaintances. However, the names used in insulting speech – so-called name-calling – are crafted using metaphors, stereotypes, and metonymies to foster greater distance from a target, not intimacy.

- (30) Thanks to the dad Pierre Trudeau & his Communist Islamic loving son of course.
- (31) Jihadi Justin always needs to be mentioned in his Post Nation as the Hero.
- (32) You think any of those women would look at you?
- (33) It's not me who looks like an idiot in this discussion.
- (34) I guess that makes you a genius by donkey standards.
- (35) You are just pretending like your life has some meaning.
- (36) You obviously don't get out much.
- (37) Think I hurt his feelings with his easy vanity degree in Literature, his public sector PC parasitic non-job, and his wife and four mistakes.

Example (30) begins with the sarcastic formula (*Thanks to*), and indirectly targets Justin Trudeau via the epithet *Communist Islamic loving son* [of Pierre Trudeau]³. Although the words “Communist” and “Islamic” are not derogatory in themselves, in the political language games that are played online they carry a highly-charged negative valence, especially when used in the construction [X]-loving. The combination of sarcasm and contextual negativity means that this clear insult is still an example of implicit offense. The alliterative nickname *Jihadi Justin* in Example (31) strengthens the connection between Trudeau and a label with a heavily negative valence. The personal attacks in Examples (32)-(36) are heightened by being so directly targeted. In (32), the target (“you”) is assumed to be an unattractive male, perhaps an “incel,” that is unworthy of the attention of the women he seems to scorn. In Culpeper’s (2016) terminology, the unpalatable question undermines the target’s self-serving assumptions about who is deserving of attention and who is deserving of scorn while, as suggested by Leech (1983), this indirectness may also sharpen the perception of impoliteness. Example (33) is a prototypical case of indirectness, and one cannot help but read it with a heavy stress on *me*. To preserve another’s feelings, it is common to say “It’s not you, it’s me.” Impoliteness, however, allows us to imply that “It’s not me, it’s you.” Still, the offense remains plausibly deniable here, since formal indirectness often decreases offensiveness. As shown in Panzeri and Carrus (2016), derogatory content can survive in conditionals and questions but tends to be diminished by indirect reports, while negated slurs are frequently rated as inoffensive.

Example (34) is reminiscent of Gore Vidal’s memorable quip about Andy Warhol, whom he described as “the only genius I’ve ever known with an IQ of 60.” Like Vidal’s quip, Example (34) has the rug-pulling structure of a classic ironic putdown. First the interlocutor is built up

³ The strict father - nurturant parent morality and the differences in worldview have an impact on understanding language, and connecting frames to certain moral categories (Lakoff 1996).

(*makes you a genius*) and then torn down (*by donkey standards*). The incongruity that is sparked by this “switch of scripts” (Raskin, 1984) or “shift of frames” (Coulson, 2001) is the source of the humor that arises from this “internal mismatch” (Culpeper, 2016) of viewpoints. The preamble to Example (35), *You are just pretending*, is unnecessary here, and is often left unsaid in an implicit offense. The remainder of the statement, *like your life has some meaning*, would work on its own as a sarcastic putdown, and the accusation of pretense weakens the wit by making the putdown less implicit than it could be. Example (36) is more formulaic as a case of sarcasm, but still requires inferencing to understand. Those who *don't get out much* may be presumed to lack the world knowledge needed to engage in serious debate, so this can be seen as another instance of seriousness silencing. Example (37) is more complex, and is characterized by heavy uses of circumlocution and figurative language. There is a note of mockery in the opening, *Think I hurt his feelings*, which implies that the target is insecure and over-sensitive. As to what they might be insecure about, the speaker goes on to dismiss, with equally mocking figurative language, the target's education, work and family. The expressions *parasitic non-job* and *four mistakes* are particularly offensive cases of dehumanization, and it is left to the reader to understand the metonymy PROCESS FOR RESULT as referring to the target's children. This accumulation of negatively valenced terms has a compounding effect. Not only are the target's children *mistakes*, they are mistakes made by a vain and lazy *parasite*.

3.2.5. Dehumanization

Dehumanization has a very specific reading in the context of implicit offense. It is not enough that a conceptual metaphor, simile or analogy causes the target to be viewed as a nonhuman entity, such as an animal or a thing, but the mapping must prompt us to think less of the target. Shakespeare did not dehumanize Juliet in this sense by comparing her to the sun, anymore than one is dehumanized by a comparison to an oil painting, an angel, or a god. Henri Bergson argued in his essays on laughter that “we laugh each time a person gives us the impression of being a thing” (Bergson 2011; Author 3 2021), and it is this inclination to laugh that is exploited by implicit offense to make a target risible in the eyes of others. Animalistic and mechanical metaphors are often effective in this regard, as are diminutives, or any exclusionary language that denies the target's humanity or agency. Dehumanization has traditionally been understood as an “extreme form of prejudice that enables violence and cruelty” (Haslam and Loughnan 2016), but recent work has revealed subtle and everyday forms of the denial of humanness.

Haslam and Loughnan (2016) distinguish two forms of dehumanization: one in which people are denied uniquely human attributes, as when they are likened to animals to imply a lack of refinement, self-control, intelligence and rationality, and another in which they are denied equally essential human attributes such as free will, as when they are likened to machines or tools. We define dehumanization as any conceptualization of a human as a nonhuman entity for purposes of causing offense. The metaphors we use – especially those used unconsciously or non-deliberately – can have a significant impact on our reasoning and shape our social and political attitudes (see Grady 2017 for an overview). Dehumanizing conceptualizations can thus promote a disregard for the human rights of certain groups (Markowitz and Slovic 2020).

In the following examples from our dataset, source concepts such as *commodity*, *cesspool*, *trash*, *dog*, *flood*, and *puppet* are used to achieve dehumanization of the speaker's targets.

- (38) Trudeau and the #Liberals have the ability to stop it but choose not to as it serves their agenda of turning Canada into a third world cesspool.
- (39) Unfortunately by the time she seeks the good guys for commitment, her value is enormously lowered due to some combination of depreciating looks, a promiscuous past, alpha widowhood, and kids needing provision.'
- (40) Better yet; that she-dog. Here girl, here girl! Good girl! Here, I will give you biscuit!
- (41) You will never be anything more than a replaceable component to be put to work.
- (42) Trailer park trash in the comments still crying.
- (43) I wouldn't be surprised if the government took more taxes off our paychecks so that Trudeau could feed and house more rapeugees that are flooding our country.
- (44) Here's some food for thought, have you ever wondered if some of the malicious sock puppets that attack you were operated by a wikipedian admin who didn't want to lose credibility because of slander?

Example (38) is both dehumanizing and racist. The literal content of a *cesspool* is not people but toxic waste, but *third world* implicitly equates this waste – via the metonymy PLACE FOR INHABITANTS – with the inhabitants of economically underdeveloped countries that are mostly non-white. Example (39) is another personalized third-person negative reference in the terms of Culpeper (2016), but its discriminatory metaphors are sexist rather than racist. A female target is objectified as a *depreciating* commodity whose *promiscuous past* has placed a burden on the state in the form of *kids needing provision*. Example (40) employs an extreme form of dehumanization to conceptualize its female target as a biddable dog that can be enticed with a biscuit. The term *she-dog* seems to be a circumlocution that avoids (yet still primes) the more

overt sexism of *bitch*. By placing its target in the submissive position of a dog eager for treats, the speaker projects themselves into a position of power and dominance. The pretense has a performative aspect which fills out what might otherwise be a partial metaphorical mapping (Lakoff and Johnson 1980). By literally acting out their metaphor (*Here girl, here girl! Good girl!*), the speaker is using what we will henceforth label a *metaphor literalization strategy*.

Examples (41) and (42) do not employ this strategy, but rather use dehumanizing metaphors for simple name-calling. Example (41) uses a mechanization metaphor that views its target as a replaceable cog in a machine. The offense is heightened with the condescending claim that *You will never be anything more*, and the speaker's sense of superiority is heightened by the assumption that they are somehow raging against the same machine, making this an example of both dehumanization and discrediting/condescending speech. Example (42) also implies the social superiority of the speaker by labeling its targets as *Trailer park trash*, which denotes the class of poor, low-class people (so-called "trailer trash") who live in trailer parks. While these people are often white (so-called "white trash"), the metaphor is not so much racist as it is classist, and the offense is compounded by suggesting its targets are not just unsophisticated but over-sensitive and baby-like (*still crying*).

Example (43) employs the neologism *rapeugees*. This is ambiguous, insofar as it might refer to the victims of rape or to the perpetrators of rape, but the larger metaphor of unwanted newcomers *flooding our country* plays to a common trope: that they are unwanted because they are criminals. It is commonplace for demagogues to stereotype a whole group based on a small number of very negative and highly salient examples (Lakoff 1987). To characterize them as perpetrators rather than victims betrays a lack of empathy and compassion, while the metaphor *flooding* further dehumanizes them as a mindless force that lacks individual agency. Example (44) also denies agency to some of its targets by using the metaphor of *sock puppets*. The example is conspiratorial in tone, but the metaphor itself is fascinating. Sock puppets are a special kind of puppet insofar as they are primitive (a hand in a sock) and unconvincing. It is their ease of detection that makes the offense implicit here, since the interlocutor is presumed to lack the basic intelligence needed to detect such a crude form of deception from others.

3.2.6. Derogation

Dehumanization diminishes its targets by treating them as non-human entities deserving of scorn, but it is possible to achieve the same end without denying their humanity. Derogatory speech, or derogation, diminishes the target by attacking the kind of human that they seem to

be. A common derogation strategy is to view the target as a child, or as a person of unusual sensitivity, or one with special needs or physical or mental impairments. The implication that the target is under-developed as a human being can be achieved with adjectives and verbs that imply incompetence, dependency, or frailty, or with rhetorical questions that challenge the target's abilities and worth. In short, dehumanization views its targets as non-human, while derogation views them as sub-human. In the following examples, the target is variously seen as a child (examples 47-53) or as a person with an illness or disability (examples 52-55).

- (45) Atta boy!
- (46) Good boys!
- (47) Did the money your grandparents sent you in a singing card by the suped up fucktarded laptop you're rocking?
- (48) I like when the little guys stick up for themselves.
- (49) Snotty 17-year-olds projecting their daddy and mammy issues on the world.
- (50) 2 questions for you, my dude: Who hurt you, and how old are you?
- (51) Grow up, sometimes things that hurt your feelings get posted to the internet.
- (52) A person with a mental illness just got elected to a position where he has a direct influence over the education of children.
- (53) But continue to lick windows at your own pleasure.
- (54) That person has nothing to offer the school board other than the fact he lives a delusional life and wants to be a 'role model' to other children.
- (55) Anyone who thinks Trudeau has done anything to support women is delusional and engaging in some next-level cognitive dissonance.
- (56) Oh tut tut, such unproven accusations are so uncivil, but posturing from a nobody pretending to be a somebody still leaves a nobody.

Examples (45) and (46) are performative in much the same way as Example (40). They literally treat the interlocutor as a child by acting out the consequences of this literal truth, and in doing so they intensify the derogatory offense. Examples (47) through (51) use a similarly derogatory conceptual metaphor, *INGÉNUE IS A CHILD*, but they differ greatly in the strategies that they employ. Example (47), the most complex, takes the form of an unpalatable question (Culpeper, 2016), and also makes extensive use of circumlocution. The metaphorical source domain of childhood is richly evoked to suggest arrested development, by imagining that the target still receives money from their grandparents, and *in a singing card* no less. Although the target is said to be *rocking* their laptop, in what is ostensibly a compliment, the derogatory neologism *fucktarded* betrays the pretense at the heart of this ironic praise. As in examples (40), (45) and

(46), this pretense is part of a larger performative act, one that fleshes out a more complete mapping between the target and a stereotypical child. Examples (52) through (56) each invoke the conceptual metaphor IGNORANCE IS A MENTAL ILLNESS to portray the target as mentally lacking. Example (52) uses the vagueness of the term *mental illness* to hint at the most damning diagnosis, while (53) uses a derogatory stereotype (*lick windows*) that is more often heard in the school yard. Example (54) uses scare quotes to reject a positive view of its target (*'role model'*) in favor of a derogatory one (*delusional life*), while Example (55) accuses its target of actively engaging in delusional behavior. Example (56) is performative in its patronizing use of *tut tut*, while *nobody pretending to be a somebody* is also a clear attempt at seriousness silencing, one that seeks to discredit the target and diminish their credibility.

3.2.8. Discrimination (racism, sexism, ageism, etc.)

Discrimination is the unjust or prejudicial treatment of individuals or groups based on their personal characteristics, such as race, gender, religion, or sexual orientation. Discriminatory language achieves these ends with various linguistic devices, including slurs, stereotypes, and exclusionary labels. Such language, which can perpetuate harmful attitudes, reinforce power imbalances, and lead to social exclusion and violence, can be explicit or implicit, but we focus here on implicit cases. Even implicit discriminatory language use may be rooted in systemic and structural forms of discrimination, such as racism, sexism, homophobia, or ableism, and can target marginalized groups by dismissing or undervaluing their particular perspectives.

- (57) Although, you are a Christian, so what more can I expect.
- (58) Wanting to live in a segregated tax-payer subsidized Islamic enclave after you immigrate to a multicultural pluralist society isn't culturally appropriate for Canada, so it isn't the right setting to do so.
- (59) What I do know is that I live on a small Island and it seems that a lot of people from various countries in the world are willing to desert their homelands and their families just to come here, in my mind that tells me a lot about said people and their home countries, most of all it highlights their weaknesses of how their own countries have let them down and how they feel the need to feed off somebody else's country, how should we English interpret this without sounding rude ???
- (60) I dunno if you're actually saying it should be that way, or just saying, it would be ideal, in a perfect world if those people didn't have kids.
- (61) Intelligent black teen does not like blacks.

(62) Ever notice that the more intelligent a person is the more they do not like blacks?

Example (57) combines the insulting speech of an *ad hominem* attack with discrimination based on religious affiliation. It is also a case of discrediting/condescending speech, since it questions the target's qualifications to express a valid opinion. Example (58), which contrasts *a multicultural pluralist society* with *a segregated tax-payer subsidized Islamic enclave*, is discriminatory in a more subtle fashion. The words comprising the latter may not exhibit an obviously negative valence, but they do appear negative when juxtaposed with the obvious positivity of the former (consider *society* vs. *enclave*, *pluralist* vs. *segregated*). Since *Islamic* is placed in the latter cluster, it too takes on a relatively negative valence. The result, which is made explicit with the clause beginning *so*, is intended to “other” and exclude Muslims. Example (59) is grounded in *white grievance* (ElSherif et al. 2021) and what Saul (2018: 365) identifies as a dog whistle working in tandem with racial resentment. The implicitness of the offense is served by the speaker's attempts to appear reasonable, but this only makes their derogatory tropes about immigrants – such as *need to feed off*, which suggests parasitism – all the more offensive. Example (60) uses circumlocution to evoke a trope about over-breeding among minorities, and equates a world with eugenics to *a perfect world*. Examples (61) and (62) also make use of implicit racism. The first licenses this racism by appearing to agree with the smartest members of the targeted group, while the second generalizes this reasoning to the *more intelligent people* of all races. The offense in the latter is sharpened by the implication that the targeted group suffers from self-hate, making other-hate acceptable, or even logical.

4. The Subtle Knife

Why do speakers resort to implicit offense when explicit offense is so much more direct, and so much easier for an interlocutor to understand? A language like English gives us a surfeit of ways to express an offensive meaning openly and without ambiguity. However, we have seen in our data analyses that implicitness holds many attractions for a speaker who aims to offend. First, implicitness affords plausible deniability to a speaker, who may later have good reason to resile from their aggressive stance. When a speaker shares a view with others who are more extreme and less subtle in their articulation of a grievance, they may employ “dog whistles” to make their position more palatable to a broader audience. Recall that a speaker using social media is addressing more than a single interlocutor and a specific target, but is simultaneously playing to the gallery too. Implicitness that does not rely on formulaic constructions demands

more creativity of a speaker, but it also affords more opportunities for play, for grandstanding, and for more eye-catching humor. A mocking linguistic turn is more cutting if its target feels that not just the speaker, but large parts of the audience, are laughing at them also. Explicit offense that relies on vulgarity and obvious epithets can be dramatic, but it rarely encourages linguistic play. In contrast, implicit offense that exploits metaphors, similes, tropes and other figurative gambits establishes a stage on which different kinds of play, from performative play acting and verbal showmanship to aggressive back-and-forth jousting, can take place.

This poses a certain danger for a speaker who finds that the interlocutor is better able to wield the figurative weapons that they have chosen for themselves. Consider a mode of figuratively adversarial humor that Author 3 et al. (2006) call “trumping.” Examples (40), (45) and (46) show how speakers can performatively treat their own metaphors as literal truths, to playfully twist the knife they have inserted into their targets. But this *metaphor literalization strategy*, as we name it here, is also available to the interlocutor, who may use the strategy to turn the metaphor against its user. It seems apt that a noteworthy user of the trumping strategy is Stormy Daniels, an adult film star who is a frequent target of former US president Donald Trump. It is alleged that after a brief affair with Daniels, Trump paid the actress for her silence and reported the payment as a legal expense. Trump denies the charges and frequently mocks the actress as “horse-face.” She, in turn, has taken ownership of Trump’s insult, and advertises her live events as “straight from the horse’s mouth.” Dynel and Poppi (2020) conducted an analysis of the linguistic strategies used by Daniels on her Twitter account to fend off abusive messages, many of which originate from Trump supporters. Their analysis shows how Daniels deftly turns abusive metaphors against their users, often by appealing to an obvious sexual meaning that has eluded her would-be tormentors. For example, when told she should “do a 180 before it’s too late” she replies “I do 180s all the time ... cowgirl to reverse cowgirl!” A recent attack by Trump on her looks inspired a supporter to tweet at the actress that “Trump wouldn’t touch you with a 10-foot pole.” She retorted by taking ownership of the figurative pole and turning it against its bearer and his master: “True. He used a 3-inch one.”

Trumping shows that metaphors are unstable, two-edged swords when one aims for implicit offense, so it is important to separate the form of the offense from the device that enables it. A question that has motivated us throughout this paper is whether modern NLP technologies can detect and grasp the full richness of the ways these devices are exploited in our data. The state of the art has made great strides in recent times with the advent of large language models, so let us consider what one such model, ChatGPT, makes of a creative instance of trumping:

User: I told my boss that "I do the work of two people for this company!" And he agreed with me, saying "Yes, Laurel and Hardy." Was my boss being complimentary or offensive?

ChatGPT: It's difficult to say for certain without knowing your boss and the context of the conversation, but it's possible that your boss was using a form of trumping that could be interpreted as both complimentary and offensive.

On one hand, your boss may have been acknowledging that you're doing a lot of work by comparing you to two people instead of one. This could be seen as a compliment, as it suggests that you're working hard and doing a good job.

On the other hand, the reference to Laurel and Hardy could be interpreted as a subtle insult. Laurel and Hardy were a comedy duo known for their bumbling antics, so your boss may have been implying that you're not as competent as you think you are. Alternatively, your boss may have been using humor to diffuse the situation and acknowledge that you're doing a lot of work without actually commenting on your competence.

ChatGPT's response is quite impressive. It starts by diagnosing our example – a confrontation between a worker and their employer – as a case of trumping. In an earlier interaction, we had defined trumping for the model as a linguistic strategy in which “one speaker undermines an idiom or metaphor used by another speaker by treating it as a literal statement.” ChatGPT now uses the term fluently, to diagnose the implicit offense at the heart of the interaction. It sees how the speaker's idiom, *the work of two people*, has been reinterpreted by the interlocutor so that it no longer signifies twice the average productivity, but now means *the work of two fools*. But it also appreciates the subtlety that comes from implicitness, noting that the reference to the bumbling duo may also serve other, more pragmatic and less biting conversational goals.

ChatGPT also has an ear for sexual euphemism; it is trained on internet content, after all. Here is how we paraphrased the Stormy Daniels “pole” interaction for the model's consideration:

User: A politician has an affair with an adult film star, but then denies it. He says "I wouldn't touch her with a 10-foot pole." She replies "True. You used a 3-inch one." What does she mean?

The model sees the interaction for what it is: an idiom-driven denial, and a sarcastic retort that plays on the meaning of “pole.” It is even shrewd enough to place this word in scare quotes:

ChatGPT: The adult film star is implying that the politician did have a sexual relationship with her, despite his denial. When the politician says "I wouldn't touch her with a 10-foot pole," he is trying to distance himself from the adult film star and suggest that he would never be involved with someone like her. However, the film star's response suggests that the politician did have a sexual relationship with her, but used a much smaller "pole" than

he claims. In other words, he did have a sexual encounter with her, but it was much shorter and less impressive than he wants people to believe.

While ChatGPT seems well-attuned to the nuances of implicit offense, it sometimes fails to appreciate the speaker's larger figurative ploy. Consider how it handles the following prompt:

User: The journalist Christopher Hitchens once said this about a controversial figure: "If you gave Jerry Falwell an enema, you could bury his remains in a matchbox." Was he being offensive? And if so, how?

Hitchens' quip is analyzed in Author 3 (2012) as a complex, highly-figurative circumlocution whose offense is heightened by its implicitness. Here is how ChatGPT understands the quip:

ChatGPT: Yes, Christopher Hitchens' comment about Jerry Falwell was intended to be offensive. The statement is a form of hyperbolic insult, which exaggerates the negative qualities of the person being criticized in an extreme and provocative way.

The use of the enema metaphor is particularly crude and graphic, and is intended to paint Falwell in an extremely negative light. The implication is that Falwell's views and actions are so odious and repellent that they could be reduced to waste matter that could fit in a matchbox.

The statement is also intended to be humorous, in a dark and satirical way. By reducing Falwell to such a base and ridiculous image, Hitchens is mocking him and undermining his credibility and influence.

ChatGPT is quick to detect the satirical insult in the quip, and it does not equivocate about its meaning. Hitchens offers and expects no quarter, so there is no plausible deniability here. The model also grasps the figurative intent of *enema*, even if its circumlocutionary function masks much of its vulgarity. Yet, while it also recognizes the hyperbolic nature of the metaphor, it fails to see the figurative significance of the *matchbox*, which is inversely proportional to its size. Ultimately, the insult is a novel variation on the explicit insult "You are full of shit." By inferring that an enema would kill Falwell and leave so little behind, we can also read it as a variation on the explicit insult "You are a total shit." By unpacking what Hitchens *could* have said, but chose not to, we appreciate his ingenuity all the more. Still, we should give credit where it is due. While ChatGPT fails to see the reason for the circumlocution, it does grasp the speaker's overall goal of discrediting Falwell by *undermining his credibility*.

5. Conclusions

In earlier research, implicitness was often explored as a matter of linguistic form, due to ease of annotation and the popularity of certain devices such as metaphor and irony. However, a focus on form alone can be misleading. While offensive similes clearly use form to be unkind, such comparisons prove extremely rare in our dataset, where they constitute just one example. Our present analysis highlights three main issues with previous typologies and datasets: (a) the conflation of content and form in modes of offense; (b) treating figurativeness, particularly metaphor, as the principal player, while ignoring this player's other leading role in explicit offense; and (c) an over-focus on formal criteria, or on specific forms that are not broadly representative of the phenomenon, such as on offensive similes.

This paper has provided an operational definition and a new typology of implicitly offensive language. The typology distinguishes between the content of an offense, where speech can be insulting, discrediting or condescending, dehumanizing, derogatory, or discriminatory, and the various forms used to express it, including metaphor, metonymy, simile, irony, hyperbole, euphemism, repetition, rhetorical questions, circumlocution, name-calling, generalization, contrastive statements, and the use of graphic devices such as all-caps text. As we have seen, many examples combine several of these devices to heighten their offensive impact, as when metaphor, irony, circumlocution and euphemism all work together to deliver a brutal payload. When an audience has to expend substantial cognitive effort to unpack an implicit offense, the effort draws others in, as with a joke, and makes them complicit in the offense (Author 3, 2012). Some of this effort is needed to negotiate the conceptual devices that are used, and some is needed to grasp the pragmatics of their use, such as whether the speaker wants us to see their target as “sub-human” or even non-human, or whether they want to silence them for their views.

Explicit and implicit offensiveness are related but ultimately very different language games. In the former, obviousness is a virtue; in the latter, it is a vice. The former wears its offense on its sleeve, while the latter makes us work to perceive it, often by masking negative intent with words that have little or no negative valence in the mental lexicon. Large language models, or LLMs, can learn to play these language games well if they are trained on very large datasets. Mechanisms of self-attention allow LLMs to focus on different aspects of a word's meaning in different contexts, and so learn how to use these words as game pieces in different games. We have seen that public-facing LLMs such as ChatGPT are already rather good at grasping the offense implicit in the mischievous use of idioms, metaphors and even sarcasm and irony. These models need no further bouts of supervised learning over large datasets, as they come *pretrained* – this is the “P” in GPT (Radford et al. 2019) – on a wealth of contextual language use. Which

is just as well, given the note of caution struck in (Wiegand et al. 2021: 584): “although generic supervised methods may always represent a good baseline, the community should also be open that other more linguistically informed approaches can be more effective for particular subtasks in the detection of implicitly abusive language.” What is needed now is not a vast dataset with annotations painted in broad strokes, but smaller, targeted datasets that are annotated in a much finer and discriminating fashion, to support the fine-tuning of LLMs. Large pretrained language models do not need copious amounts of additional training data, but can learn from a few targeted examples, in what is called *few-shot learning* (Brown et al., 2020). Anyone with a web browser can see for themselves how far LLMs have already come. It will take finesse and insight, not scale and generality, to take them the rest of the way.

References

Author 1 et al. 2019.

Author 3 et al. 2006.

Author 3. 2012.

Author 3. 2021.

Barnden, J. 2016. Communicating Flexibly with Metaphor: A Complex of Strengthening, Elaboration, Replacement, Compounding and Unrealism. *Review of Cognitive Linguistics* 14 (2): 442-473.

Bergson, H. 1911. *Laughter: An Essay on the Meaning of the Comic*. Translated by Cloudesley Brereton and Fred Rothwell. New York, NY: The Macmillan company.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th NeurIPS Conference on Neural Information Processing Systems*, Vancouver, Canada, volume 33, pages 1877– 1901.

Caponetto, L. 2021. A Comprehensive Definition of Illocutionary Silencing. *Topoi* 40: 191– 202.

Caselli, T.; Basile, V.; Mitrović, J.; Kartoziya, I.; Granitzer, M. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille: European Language Resources Association (ELRA), pp. 6193–6202.

Coulson, S. 2001. *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. Cambridge, UK: Cambridge University Press.

Culpeper, J. 2011. *Impoliteness: Using Language to Cause Offense*. Cambridge: Cambridge University Press

Culpeper, J. 2016. Impoliteness Strategies. In: Capone, A., Mey, J. (eds) *Interdisciplinary Studies in Pragmatics, Culture and Society. Perspectives in Pragmatics, Philosophy & Psychology*, vol 4. Springer, Cham. https://doi.org/10.1007/978-3-319-12616-6_16

Dancygier, B., & Sweetser, E. 2014. *Figurative language*. Cambridge University Press.

Dynel, M. & Poppi, F.I.M. 2020. Arcana Imperii: The power of humorous retorts to insults on Twitter. *Journal of Language Aggression and Conflict* 8 (1), 57-87.

ElSherief, M.; Ziems, C.; Muchlinskiy, D.; Anupindi, V.; Seybolty, J.; De Choudhury, M.; Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 345–363.

Gao, L.; Kuppersmith, A.; Huang, R. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*. Taipei, Taiwan: AFNLP, pp. 774–782.

Gibbs, R. W.; Colston H. L. 2012. *Interpreting Figurative Meaning*. New York: Cambridge University Press.

Grady, J. 2017. Using metaphor to influence public perceptions and policy: Or, how metaphors can save the world. *The Routledge Handbook of Metaphor and Language*. Routledge. Eds. Semino, Elena; Demjén, Zsófia. Routledge. Taylor and Francis Group. London – New York.

Grice, P. 1975. Logic and Conversation. Pp. 41–58 in *Syntax and Semantics 3: Speech Acts*, edited by P. Cole and J. J. Morgan. New York, NY: Academic Press.

Haslam, N., Loughnan, S., Reynolds, C., & Wilson, S. 2007. Dehumanization: A new perspective. *Social and Personality Psychology Compass*, 1(1), 409–422. <https://doi.org/10.1111/j.1751-9004.2007.00030>.

Kumar, R.; Ojha, Atul Kr.; Malmasi, S.; Zampieri, M. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC)*. Santa Fe, pp. 1–11.

Kunupudi, D.; Godbole, S.; Kumar, P.; Pia, S. 2020. Toxic language detection using robust filters. *SMU Data Science Review* 3(2). Available at: <https://scholar.smu.edu/datasciencereview/vol3/iss2/12>.

Landau, M. J., Keefer, L. A. and Swanson, T. J. 2017. ‘Undoing’ a Rhetorical Metaphor: Testing the Metaphor Extension Strategy. *Metaphor and Symbol* 32 (2): 63-83. DOI: 10.1080/10926488.2017.1297619

Leiter, C.; Zhang, R.; Chen, Y.; Belouadi, J.; Larionov, D.; Fresen, V.; Eger, S. 2023. ChatGPT: A Meta-Analysis after 2.5 Months. *ArXiv* 2302.13795.

Li, H. 2022. Language models: past, present, and future. *Communications of the ACM*, volume 65, number 7, pages 56–63.

Markowitz, D. M.; and Slovic, P. 2020. Social, psychological, and demographic characteristics of dehumanization toward immigrants. *PNAS* 117 (17), 9260-9269

Myers, D. G. 2013. *Social psychology*. Twenge, Jean M., 1971- (11th ed.). New York, NY: McGraw-Hill. ISBN 978-0-07-803529-6. OCLC 795645100.

Panzeri, F., & Carrus, S. 2016. Slurs and negation. *Phenomenology and Mind* 11, 170-180. <https://tinyurl.com/36trde54> Accessed 27 Feb. 2023.

Patrick, C. and Hollis, L. 2018. Contentious Cloud Chatter: A Comparative Analysis of Aggressive Speech. *Comparative Civilizations Review* 78 (8). Available at: <https://scholarsarchive.byu.edu/ccr/vol78/iss78/8>

Qadir, A.; Ellen Riloff, & Walker, Marilyn A. 2015. Learning to Recognize Affective Polarity in Similes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–200, Lisbon, Portugal.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* volume 1, number 8, page 9.

Raskin, V. 1984. *Semantic Mechanisms of Humor*. Dordrecht, Netherlands: D. Reidel.

Sadeque, F.; Rains, S.; Shmargad, Y.; Kenski, K.; Coe, K. & Bethard, S. 2019. Incivility Detection in Online Comments. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. 283–291. Minneapolis, Minnesota. Association for Computational Linguistics.

Saul, J. M. 2018. Dog whistles, Political Manipulation, and Philosophy of Language, in Daniel W. Harris Daniel Fogal, and Matt Moss (eds.), *New Works on Speech Acts*, Oxford: Oxford University Press, pp. 360–383.

Schmidt, A.; Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics. Valencia. 1–10.

Searle, J. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511609213

Stoll, A.; Ziegele, M.; Quiring, O. 2020. Detecting impoliteness and incivility in online discussions. Classification approaches for German user comments. *Computational Communication Research* 2(1). 109–134. DOI: 10.5117/CCR2020.1.005.KATH.

Van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 33–42, Brussels, Belgium.

Waseem, Z., Davidson, T., Warmusley, D., & Weber, I. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online* (pp. 78–84). Vancouver. Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W17-3012> doi: 10.18653/v1/W17-3012.

Wiegand, M.; Geulig, M.; Ruppenhofer, J. 2021b. Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis. In Merlo, P.; Tiedemann, J.; Tsarfaty, R. (Eds). *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021*, 358–368.

Wiegand, M.; Ruppenhofer, J.; Eder, E. 2021a. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (Eds.): *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: Association for Computational Linguistics, pp. 576–587.