Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts

Luke M. Breitfeller[♠] **Emily Ahn**[♠] **David Jurgens**[♢] **Yulia Tsvetkov**[♠]
[♠]Carnegie Mellon University [♢]University of Michigan
{mbreitfe, eahn1, ytsvetko}@cs.cmu.edu, jurgens@umich.edu

Abstract

Microaggressions are subtle, often veiled, manifestations of human biases. These uncivil interactions can have a powerful negative impact on people by marginalizing minorities and disadvantaged groups. The linguistic subtlety of microaggressions in communication has made it difficult for researchers to analyze their exact nature, and to quantify and extract microaggressions automatically. Specifically, the lack of a corpus of real-world microaggressions and well-defined criteria for annotating them have prevented researchers from addressing these problems at scale. In this paper, we devise a general but nuanced, computationally operationalizable typology of microaggressions based on a small subset of microaggression data that we have. We then create two datasets: one with examples of diverse types of microaggressions recollected by their targets, and another with gender-based microaggressions in public conversations on social media. We introduce a new, more objective criterion for annotation and an active-learning based procedure that increases the likelihood of surfacing posts containing microaggressions. Finally, we analyze the trends that emerge from these new datasets.

1 Introduction

Toxicity and offensiveness are not always expressed with toxic language. While a substantial community effort has rightfully focused on identifying, preventing, and mitigating overtly toxic, profane, and hateful language (Schmidt and Wiegand, 2017), offensiveness spans a far larger spectrum that includes comments with more implicit and subtle signals that are no less offensive (Jurgens et al., 2019). One significant class of subtle-but-offensive comments includes *microaggressions* (Sue et al., 2007, MAs), defined in Merriam-Webster as "a comment or action that subtly or often unconsciously expresses a prejudiced attitude toward a member of a marginalized group such as a racial minority." Though subtle, MAs have been shown to have lasting harmful impacts on their targets. Qualitative interviews suggest that the subtlety of MAs may cause even greater levels of situational stress than overt aggression (Sue, 2010; Nadal et al., 2014).

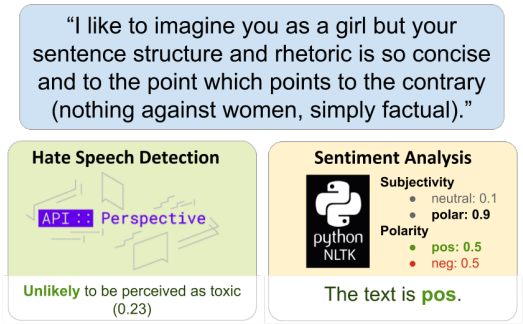

Figure 1: Existing state-of-the-art tools for hate speech detection and sentiment analysis cannot identify the veiled offensiveness of microaggressions (MAs) like the one in this real comment, because in many cases, the framing of a MA includes stylistic markers of positive language. This motivates the need for a corpus and new tools for detecting MAs.

Despite a public effort to recognize and reduce—if not eliminate—their occurrence (Kim, 2013; Neff, 2015), there has been no computational work to detect and *analyze* MAs *at scale*. Instead, much of the recent work has focused on explicitly toxic language (e.g., Waseem et al., 2017), with surveys of the area also overlooking this important and challenging task of recognizing this subtle toxicity (van Aken et al., 2018; Salminen et al., 2018; Fortuna and Nunes, 2018). Indeed, as Figure 1 suggests, current popular tools for toxic language detection do not recognize the toxicity of MAs and further, sentiment tools can label these comments as being positive. As a result, applications using such techniques to promote inoffensive content may potentially *promote* MAs. Such biased online content is then used

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1664–1674, Hong Kong, China, November 3–7, 2019.

as training data in NLP tools—dialogue systems, question answering, and others—thereby perpetuating and amplifying biases (Zhao et al., 2017).

Computational modeling of MAs is challenging for many reasons. MAs are subjective, context-sensitive, and expressed subtly in language (as shown in Figure 1); there are no well-defined annotation guidelines, no corpus of MAs for training and evaluating a model, and prior computational approaches to detecting overtly offensive speech are likely not suitable for classifying MAs. Moreover, they are diluted in the ocean of social media content, and it would be infeasible to create a corpus of diverse types of MAs just by crowdsourcing annotations of randomly sampled social media posts. This paper makes a first step towards closing this gap, providing (i) a typology of MAs, (ii) a corpus, and (iii) a first computational model to surfacing MAs.

Here, we introduce the first computational analysis of MAs with three main contributions. First, we create a seed corpus through collecting self-reported accounts of MAs online; we analyze this data using prior research in social science on MAs in order to derive an operationalizable classification scheme (§2.2), which we show can be used as training data to classify types of MAs (§2.4). Second, we develop a new active-learning based crowdsourcing technique to identify candidate microaggression messages for annotation (§3). In a large-scale study of gender-based MAs on Reddit, we show that our method is effective at surfacing MAs. Our novel approach surfaces MAs not found through simply looking for offensive posts (§3.3). Finally, we compare the types of MAs from self-reported accounts and those observed in social media comments, and we draw new conclusions about how MAs behave and may be identified in online settings (§4). [1]

2 A Corpus of Microaggression Classes

Previous examinations of MAs have typically focused on small scale studies with interviews, surveys, or discussions with clinicians (e.g., Shelton and Delgado-Romero, 2013; Campbell and Manning, 2014; Forrest-Bank and Jenson, 2015), aiming at understanding the causes and effects of MAs on their targets (Wong et al., 2014). A crucial gap in existing MAs research is the lack of systematic analysis of reports (Lilienfeld, 2017), which has contributed to the difficulty of understanding what constitutes a microaggression. Here, we aim to address this gap by introducing a new dataset of MAs and then use this data and prior qualitative research to construct a new typology of MAs. This new typology enables us to ground the notion of what is a microagression, which also addresses criticism that the concept is too slippery. The typology is critical for downstream analyses of determining who is affected and how they are being targeted. Further, we use the typology as a practical heuristic for identifying which types of MAs are the highest impact for computational approaches. As a test, we develop a classifier using the typology and show that different types of MAs are readily distinguished.

2.1 SELFMA: microaggressions.com Dataset

The Tumblr website www.microaggressions.com curates a collection of self-reported accounts of MAs. They were collected by asking people to fill out an online form with three questions: the story of the microaggression, the context, and how the microaggression made them feel. As such, the microaggression accounts in this corpus come from individuals who self-report the incident, and who, presumably, are at least familiar with the term "microaggressions."

From this website, we collected 2,934 accounts of MAs (henceforth, *posts*). Most posts are manifestations of bias, targeting social groups frequently discriminated against, including gender (1,314 posts), race (1,278 posts), sexuality (461 posts), and religion (88 posts), among others. Some are formatted to show the conversation structure of the recollection of the incidents, while some others are shown as plain narratives. We refer to this dataset as SELFMA.

2.2 A New Typology of Microaggressions

As originally specified in Sue et al. (2007), MAs were categorized into 13 themes across specific axes of discrimination such as racism and sexism. These themes were developed qualitatively from transcripts with therapists. However, while they were comprehensive in the clinical domain, here, we aim to create a more general framework for classifying MAs in online interactions. Specifically, our new typology builds upon the work of Sue et al. (2007) with three goals in mind: the ty-

[1] All data and code produced through this work is available at https://bit.ly/2lVv3BG. This study was approved by the Institutional Review Board (IRB).

pology should be (1) generalizable across different axes of discrimination, (2) sufficiently represented in our corpus, and (3) comprehensive over all distinct microaggression types in the corpus.

Four key themes were identified in our analysis of the MAS data: Attributive, Institutionalized, Teaming, and Othering (see Table 1). We discuss each of these next.[2]

The **Attributive** theme covers instances where a microaggression attributes a stereotype to an individual based on their identity. These stereotypes may have inherently negative connotations ("lazy"), but may also be neutral ("liking pink") or positive ("strong"), which complements recent work on benevolent sexism (Jha and Mamidi, 2017). The **Institutionalized** theme reflects larger institutionalized biases, such as in employment or law enforcement. The **Teaming** theme is derived from the term *forced teaming*, coined by de Becker (1997) to describe a strategy of abuse where the abusers frames themselves as being on the same team as the victim. The **Othering** theme covers MAS which revolve around framing the target in relation to some "othered" group. This theme often co-exists with "Attribution of Stereotype," but is distinct in that its focus is on redefining the target's sense of identity.

2.3 Annotation Results

Three annotators familiar with the theoretical background and prior research on microagressions performed an open coding procedure after examining the SELFMA data to codify the typology and determine annotation guidelines. All three labeled 200 instances of the dataset to estimate agreement.

Despite the difficulty of the task, annotators had moderate agreement, as shown in Table 2. While lower than what is considered high agreement (Artstein and Poesio, 2008), given the potentially-subjective nature of MAS and criticism for lack of objectivity (Lilienfeld, 2017), we view this result as a strongly positive sign that reliable annotation is possible despite the challenge. This moderate agreement is on-par with other difficult annotation tasks, such as those for connotation frames (Rashkin et al., 2016), which had 0.52 percentage agreement for rating the polarity of the sentence towards a target, dimensions of social relationships (Rashid and Blanco, 2017), which had κ values as low as 0.59, and Word Sense Disambiguation (Passonneau et al., 2012), which reported α values for some words below 0.30 at determining meaning.

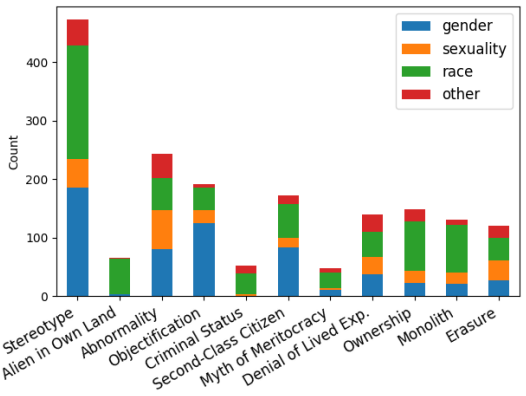

Figure 2: The distribution of four axes of discrimination (gender, race, sexuality, and other) in each subtheme.

The final dataset was determined by first retaining all posts where at least two annotators agreed (183 posts). And where there were no agreement (17 posts), the annotators determine the labels through a follow-up adjudication process.[3] After this process, another 1,100 posts were singly annotated, distributed across the three annotators, for a total of 1,300 posts. Relative percentages of each theme and examples comments are shown in Table 3. We also show the distribution of different axes of discrimination among the subthemes in Figure 2.

2.4 Classifying Types of Microaggressions

As a further examination of the typology, we perform a study where a model is trained to classify a comment into one of the four themes. Linear SVM classifiers were trained in a one-versus-rest setup using (i) word unigrams and bigrams on the tokenized version of the posts,[4] (ii) categories in the 2015 LIWC lexicon (Pennebaker et al., 2015), (iii) sentiment lexicons, and (iv) the topic distribution for a 40-topic LDA model (Blei et al., 2003). They were trained on 1,000 posts and evaluated using a held-out test set of 300 posts.

Classifier performance, measured as F1, was on

[2]The full description of themes, sub-themes, and examples can be found in Supplementary Material §2.

[3]There were 26 posts deemed to be exclusively overt aggressions (which could be detected using existing hate speech classifiers), and these occurred only in the training set. We removed them from our corpus.

[4]We used NLTK `wordpunct_tokenize` on the concatenated texts from the structured and unstructured format of the posts, normalizing two or more repeated characters into two. We only take those unigram and bigram features which appear in at least two posts as our feature set. This gives us 8,853 distinct features in the training data.

	Sub-theme	Definition and Example
Attributive	Attribution of Stereotype	Link some attributes to an individual based on their identity. "Girls just aren't good at math."
	Alien in Own Land	Marginalized individuals are foreign. "But where are you from, originally?"
	Abnormality	Marginalized individuals are abnormal. "Why do we need the word cisgender? That's just normal people."
Institutionalized	Objectification	Diminish the humanity of marginalized individuals. "If you don't want to get hit on, wear a longer skirt."
	Criminal Status	Link a persons identity to criminality, danger, or illness. "You look like a terrorist with that beard."
	Second-Class Citizen	Marginalized individuals belong to low-status positions in society. "Oh, you work at an office? I bet you're a secretary."
Forced Teaming	Myth of Meritocracy	Differences in treatment are due to ones merit. "They just cast actors who are best for it. Why does it matter if they're all white?"
	Denial of Lived Experience	Minimize the experiences of a marginalized individuals. "It was just a joke! You're too sensitive."
	Ownership	Anyone can have some claim to a marginalized groups experiences. "Why is it offensive for a white person to wear a bindi? It's just jewelry."
Othering	Monolith	All members of a marginalized group are identical. "My gay friend doesn't have a problem with this show. I don't get why you're mad."
	Erasure	Anyone can claim that an individual does not belong to that group. "Your mom is white, so it's not like you're really black, though."

Table 1: Sub-themes that we developed based on the data from www.microaggressions.com, and short excerpts that show the main claim or assumption of each sub-theme. Examples demonstrate that MAs can harm and invalidate in a subtle way, typically targeting disadvantaged populations. The phenomenon is complex, different from overt profanity or negative sentiment.

Theme	Fleiss' κ
Attributive	0.4156
Institutionalized	0.5058
Teaming	0.5048
Othering	0.4300
Overall	0.4641

Table 2: Inter-annotator agreement across three annotators on the four themes of microaggressions.

par with annotator agreement at 44.25. However, individual themes varied substantially in their performance, indicating not all were easily recognizable. Attributive and Institutionalized had the highest performance at 58.33 and 44.76, respectively; in contrast, Teaming and Othering had far lower F1 scores at 32.69 and 15.87, respectively. Both Teaming and Othering were the least frequent themes in the corpus, increasing the difficulty of recognizing them from limited data.

3 Locating Microaggressions in the Wild

The vast majority of microaggression data comes from self-reported recollections, including the MAs corpus that we collected. Self-reported data is crucial for studying the phenomenon and its effects, but is arguably not suitable for training classifiers, due to two critical limitations: (i) biases in how the offending comment was originally made versus how it was recalled, and (ii) biases in which kinds of microaggression comments are recalled. These make future computational attempts at detecting MAs more difficult by skewing a model away from linguistic signals used in real-world—rather than recalled—behavior.

A natural solution is to collect and annotate examples of MAs, scaling this process by using crowdsourcing. There are, however, two critical obstacles to using the straightforward approach of asking crowdworkers to identify random social media posts as MAs: (1) untrained crowdworkers are unlikely to reliably identify these posts due to the potential *subjectivity* of MAs, and (2) a naive random sampling of social media posts is unlikely to uncover MAs due to their *sparsity*, and since MAs are veiled manifestations of biases, it is also not clear how to use lexicons as a sieve for a stratified selection of posts to annotate.

In this second part of the paper, we devise a novel methodology to address the two challenges to curate microaggression corpora from social media posts with a two-pronged approach: operationalizing and surfacing MAs.

Operationalizing Microaggressions As the first study to crowdsource MAs, we focus on finding examples of gender-based MAs. By focusing

Category	%	Examples	Toxicity
Attributive	50	For a female, you've got backbone.	.12
		There's fighting, for you boys, and romance, for you ladies	.10
Institutionalized	28	So you make the coffee and he does the work?	.07
		Summer's coming! Gotta fit into that tight swimsuit.	.23
Teaming	23	Most workplace harassment is an example of women being oversensitive to advances of men	.27
		So were you good enough to get in or did you get in through a women's quota?	.19
Othering	17	I don't want another woman boss. Already have one and she has me running around all the time	.17

Table 3: The final microaggression typology after analyzing 1,300 self-reported microaggressions in SELFMA dataset. As shown by the percentages of instances covered each category, this typology effectively summarizes the broad themes (note that one instance may be labeled with multiple categories). For each category, the overall toxicity score (measured in [0,1] using the Google Perspective API) is low, indicating that currently none are reliably recognized as being offensive.

on gender-based MAs, the dominant group (male annotators) and the marginalized groups (women and non-binary annotators)[5] are easily elicited from the annotators through self-reporting. Since training crowdworkers to directly identify MAs is infeasible, we look for a more objective proxy that can identify, or at least help surface, posts containing MAs. A key property of MAs is that they often result from different lived experience between the individuals in the dominant group and those in the marginalized groups. Thus, our hypothesis is: *there will be a discrepancy of perceived offensiveness between the dominant group and the marginalized groups for* MAs. If true, we can then ask annotators to annotate the perceived offensiveness of a statement, and use this as a proxy to surface MAs from social media posts. Based on Figure 2, we pick gender as the discrimination axis for this work, since it covers a large number of our posts.

Surfacing Microaggressions To alleviate the sparsity issue, we treat the annotation process as an active learning procedure, where batches of posts to be annotated are provided by a classifier that is trained to predict the discrepancy of perceived offensiveness of a post, which we have claimed above to be a proxy to MAs. This classifier can then be iteratively updated using the results of all previous batches. Assuming the classifier is good enough in predicting the proxy, the classifier will help surface the most likely candidates of posts that contain MAs. In what follows, we describe the annotation process in detail and the resulting crowdsourced dataset.

3.1 Crowdsourcing Microaggressions

The full data collection process is set-up using a *multi-round annotation scheme*, summarized as follows. We first select a random sample of Reddit posts from relevant subreddits. These posts are then given to crowdworkers to be annotated for perceived offensiveness. This first round of annotations is then used to train a classifier for a MA-relavent task. The classifier's ratings are used to pick the next batch of Reddit posts to be annotated, in order to surface more instances than random sampling.

Task Crowdworkers are shown comments in the context of an internet forum discussion. Crowdworkers are asked to imagine they have written the displayed post, and then after a button click, we reveal how someone (a replier) has commented to that post. They are then asked rate if they found the response offensive and if so, to what degree was it offensive on a seven-point Likert scale. This scale of offensiveness allows us to capture cases where a microaggression is perceived as offensive by both annotator groups (genders), but to different degrees. To test our hypothesis and quantify how certain posts could be perceived differently by annotators of different genders, each annotation task includes a demographic question.[6]

Data To focus the data on interactions where gender may be a salient variable, Reddit data was drawn from RtGender (Voigt et al., 2018), which has a post-and-reply interactions where the gender of each author has been inferred with high confidence. After preliminary analyses, the initial data was randomly selected from 28 subreddits based on their focus around gender issues (e.g,. Relationships, AskWomen).[7]

For the **first round of annotation**, we used these four subsets of posts to be annotated for of-

[5]Even this interpretation of gender is limited, as women and nonbinary individuals experience gendered discrimination in different ways, and transgender men also have a marginalized gender identity.

[6]Following best practices in gender elicitation (Jaroszewski et al., 2018), we include a third option of "nonbinary, genderqueer, or otherwise", which has a free-form text box. We opted not to ask about cis or trans status in the demographic survey.

[7]See Supplementary Material §1 for the complete list.

fensiveness by the crowdworkers:

1. (SELFMA) Posts from our MAS corpus (§2) that contain gender-based MAS ($n = 59$);
2. (NON-OFFENSIVE) Posts manually picked from the Reddit dataset that should quite clearly be perceived as not offensive by everyone ($n = 36$);
3. (OFFENSIVE) Posts manually picked from the Reddit dataset, filtering by profane lexicons targeting women, and a corpus of Twitter hate speech (Davidson et al., 2017) that should quite clearly be perceived as offensive ($n = 21$);
4. (RANDOM) Posts randomly selected from the 28 subreddits ($n = 1007$).

The first three sets of posts serve as controls both to verify the soundness of our hypothesis (SELFMA) and to filter out untrusted crowdworkers (NON-OFFENSIVE and OFFENSIVE).

Adaptive Data Selection As described previously, after the first round of annotations, we use a classifier trained to predict discrepancy in perceived offensiveness to pick the next batch of posts for the **second round of annotation**. The classifier was given, as positive examples, posts with ≥ 0.25 discrepancy[8] between the averaged perceived offensiveness of the dominant group and the marginalized groups. Other posts are considered negative examples. The feature sets that we used are motivated by prior work on gender bias and power dynamics. The following seven feature categories are used: (1) unigram and bigram features to capture lexical patterns, (2) two categories of formal and informal words derived from the formality corpus of Pavlick and Tetreault (2016), (3) the gendered occupations lexicon of Bolukbasi et al. (2016), grouped across definitional and stereotypical gender (male, female, neutral), (4) the gender stereotype lexicon of Fast et al. (2016), (5) the gender lexicon for social media from (Sap et al., 2014), (6) a manually-compiled corpus of gendered words, extended from seed list from https://www.hrc.org/resources/glossary-of-terms and (7) a manually-compiled sentiment lexicon, inspired by LIWC. To facilitate reproducibility, all lexicons will be available in the software release.

[8]The threshold was chosen empirically to ensure enough number of positive examples.

3.2 Experimental Setup

We designed the annotation task with two purposes in mind: (1) We want to test our hypothesis that there is discrepancy in how annotators of different genders perceive MAS. By comparing the distribution of discrepancy of perceived offensiveness in SELFMA, which consists solely of MAS, and of that in OFFENSIVE, which consists solely of offensive posts, we can see whether the discrepancy test actually is an **operationalizable criteria** to distinguish MAS from offensive posts in general. (2) We want to test our hypothesis that the discrepancy can be used **to surface MAS**, better than the alternatives. For the purpose of this experiment, we also trained a classifier to predict the offensiveness level of a post, and use that to provide a batch of posts to be annotated as well. By comparing the actual number of MAS by these two classifiers, we can test whether discrepancy can be used to surface more MAS compared to offensiveness in general.

Platform Setup Data was crowdsourced using Figure Eight with only Level 3 workers, a small "group of most experienced, highest accuracy contributors". Workers were shown 10 posts and replies per page, one of which was a test question. Workers whose ratings were substantially outside of the rating range of our ground truth annotations in more than 60% of cases were removed. Workers were paid $0.40 per page. To control for language and cultural norms, workers were restricted to being from the US, UK, Canada, and Australia.

Much like Mechanical Turk (Hara et al., 2019), Figure Eight has substantial representation of female-identifying individuals (Posch et al., 2018), with slightly more than 60% of users in the US identifying as female on both platforms (a higher rate than other countries). Studying gender-based MAS on these platforms means we are more likely to get a gender-balanced sample than compared with other social categories subject to microaggressions such as religion, age, or sexual orientation, which would likely require targeted recruiting of respondents or a massive collection of responses in order to achieve a balanced sample.

3.3 Results

For the first experiment, Figure 3 shows the distribution of perceived offensiveness discrepancy between the dominant group (men) and the marginalized group (women and non-binary) an-

Figure 3: Difference in perceived offensiveness between annotator genders show that MAs in the SELFMA were perceived as substantially more offensive by annotators identifying as women (shown through a positive rating discrepancy on y-axis) than offensive comments in OFFENSIVE. Difference is significant under t-test ($p < 0.001$).

notators, which supports our hypothesis that MAs in SELFMA are perceived as substantially more offensive by annotators identifying as women and non-binary.

For the second experiment, we asked crowdworkers to annotate 2,000 posts with the highest scores as predicted by the classifier, and then we took the top 6% of Reddit posts with the highest discrepancies scores and manually annotated them for MAs. We predicted that these posts contain a higher percentage of MA posts than those selected randomly, and this hypothesis is supported: Out of these aggregate 120 posts, 13 are MAs (10.8%),[9] compared to a manually annotated 200 random posts from the RANDOM batch, where only 6 of them (3%) were MAs. Examples are given in Table 4.

A potential next hypothesis suggested by our original hypothesis is that the classifiers trained on annotator discrepancy (with a higher level of offense for female and non-binary annotators) would be more successful in locating MAs than the classifiers trained on pure offensiveness. However, we did find equal performance or slight gains in microaggression location when using the offensiveness classifier. Based on analysis of the SELFMA data, we believe the cause may be that gender-based MAs were not always annotated as being more offensive to female annotators–many had strong discrepancies with male annotators being more offended. As our annotator discrepancy classifier was trained on samples more offensive to women, these types of MAs would be filtered out and decrease the model's overall effectiveness at surfacing MAs.

[9]The 2,000 posts were selected by taking the top 500 posts from the discrepancy and offensiveness classifiers on two different subsets of the Reddit data. Out of the 13 MAs, 6 come from the discrepancy classifiers, and 7 come from the offensiveness classifier.

4 Reported vs Observed Themes

How do self-reported and passively-observed MAs differ in thematic frequencies? Here, we use our thematic classifier (§2.4) to label all the microaggression data in Reddit. Analyzing the thematic differences reveals substantial differences between the two datasets. The Stereotype theme was by far the most common in our SELFMA data (53%), but it rarely occurs in our Reddit (3%). Instead, Objectification is the most common, as shown through the Institutionalized category (79%). Analyzing *where* the MAs occurred in Reddit, we observed that of the 19 instances of MAs identified by our annotators, many came from strongly male-focused subreddits (e.g., MensRights, Tinder, seduction, justneckbeardthings), where it would be expected that one male poster is talking to another. These online settings mirror offline highly-gendered settings where casually sexist language is known to happen, e.g., "locker room talk" (Katz, 1995). This is embodied in our Institutionalized category (specifically the subcategory "Objectification") having a much stronger presence in the Reddit MAs than the SELFMA ones in Table 5. While such language may not be perceived as offensive in the original dialog context, other readers who encounter it would likely feel offended and potentially marginalized.

Our work further points to two potential limitation of self-reported statements. First, the baseline rates of MAs may not be reliably estimated since the offensive comments are made elsewhere (outside of direct experience) and would only be discovered to through observational analysis. Second, the difference in baseline rates suggests that the relative differences in perceived offensiveness by theme may result in different reported quantities of microagressions; as not all comments and themes are equally offensive, people are more likely to recall the extremes of events (Tversky and Kahneman, 1973), potentially contributing to the difference in thematic rates in our two datasets.

5 Discussion and Limitations

As a first computational study of MAs, our work comes with some notable limitations that each provide clear avenues for future work. First, the current dataset contains just over 6.5K instances. While useful, this corpus does is likely not fully

Classifier	Category	Subreddit : Examples
Discrepancy	Institutionalized	r/skateboarding: "this isn't a real competitor though. She's just a gross bimbo looking for attention [...]"
Discrepancy	Teaming	r/changemyview: "[...] Don't you think you ignored the real points I was making? Puting him in a "pro"-molester group, when he maybe just be pro tolerance?"
Offensive	Institutionalized	r/Tinder: "I'd f*** a stupid girl, but I definitely won't date or marry her. Other intelligent men are the same."
Offensive	Teaming	r/MensRights: "[...] I myself don't want to get the police involved over some girl saying she'll dump a guy or have an affair if he doesn't meet her sexual standards."

Table 4: Examples of MAs surfaced by the discrepancy/offensiveness classifiers, with corresponding manually-annotated labels of their typology category.

Category	Percentage SELFMA (§2.1)	Reddit
Attributive	53	3
Institutionalized	59	79
Teaming	10	21
Othering	7	10

Table 5: Gender-based MAs crawled in SELFMA ($n = 59$) and found in Reddit ($n = 19$) after the second round of annotation, according to their category from our proposed typology (§2.2).

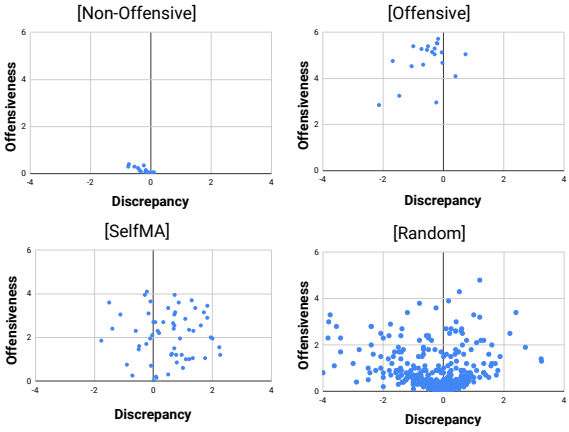

Figure 4: Offensiveness (y-axis) vs Discrepancy (x-axis) of perceived offensiveness between annotator gender. for the four sets of posts (§3.1). Note that compared to clearly offensive data [OFFENSIVE] and clearly non-offensive data [NON-OFFENSIVE], the SELFMA data (§2.1) identified as gendered MAs are annotated as somewhat offensive (indicated by the distribution around the middle of the y-axis), and having more discrepancy in perceived offensiveness between annotator genders (indicated by the higher degree of spread on the +x-axis). RANDOM data collected from Reddit sit all along the two axes of offensiveness and discrepancy.

representative of all types of MAs. However, our crowdsourcing approach does provide an effective way to surface MAs and our dual objective approach, which uses both annotator discrepancy and offensiveness, provides complementary views into what statements could be perceived as MAs. Additional iterations of this procedure are likely to improve microaggression recognition and substantially increase the seize of the corpus. We note that one option is to have workers *generate* examples, rather than rate (e.g., Xu et al., 2013; Su et al., 2016; Jiang et al., 2018); however, such a process raises ethical concerns of having crowdworkers generate toxic statements towards others.

Second, our current focus is on gender-based MAs. This choice was motivated by the observation that gender-based MAs are the largest category in the SELFMA data and, given that prior studies have shown substantial gender disparity online, with women receiving more negative behaviors (Duggan, 2017), this choice has the potential for highest impact. Our work builds upon a growing body of literature focused on identifying and mitigating gender disparity through computational means (e.g., Magno and Weber, 2014; Garimella and Mihalcea, 2016; Li et al., 2018; Field et al., 2019; Field and Tsvetkov, 2019). Further, our focus on gender also allows us to reliably recruit crowdworkers across the gender spectrum, whereas other social categories such as race or religion are more difficult to recruit in a balanced proportion through traditional mechanisms. However, despite this current focus, both the typology and annotation approach are designed for and readily amenable to other societal dimensions of MAs.

Third, performance could still be improved even for the targeted case of gender-based MAs. For example, Figure 4 plots the distribution of posts according to the averaged perceived offensiveness (y-axis) and the discrepancy between the dominant group and the marginalized groups (x-axis). This shows that offensiveness and discrepancy in conjunction are better features than using just discrepancy in separating MAs from simply offensive and non-offensive posts. A possible future direction could be to train a multi-task classifier to predict both features in order to choose posts to annotate.

A final concern is that our approach may not truly be an "objective" replacement to current MAIdentification, as annotator bias is still present in this process. By choosing perceived offensiveness and discrepancies in perception as proxy measures for MAs, we hope to shift studies of

MAs from a judgment call on an individual's intent to a description of how it affects its targets, which can be described as an objective measure of human behavior that can be validated through study. However, it is true that without that validation, it is difficult to prove the behavior of our crowdsourced annotators represents inherent, objective truths about how groups react to MAs.

6 Conclusion

Offensive language can take many forms: seemingly positive comments can in reality be dismissive of the experiences and validity of other persons. MAs, everyday comments that marginalize others on the basis of their identity, are a common form of this linguistically subtle offensive language. While prior work in NLP has largely focused on overtly offensive language, efforts to recognize this more insidious and equally impactful language has been stymied by the lack of data and methods for effectively annotating large corpora. In this work, we posit that covert toxicity can be detected through contextualized analysis of subtle linguistic cues: pronoun uses, adjectives, mentions of social groups, stylistic cues, etc. We thus view the task of detection of MAs as a yet unexplored, socially-relevant NLP task, somewhat similar to (but possibly harder than) sentiment analysis and detection of hate speech.

This work makes three key contributions. First, we introduce a new thematic classification of MAs and show that annotations can reach moderate agreement on a highly difficult task–countering claims that microaggressions are too subjective to be reliable–and also showing that computational methods can distinguish these themes. Second, we introduce a new active-learning crowdsourcing approach for surfacing potential MAs from social media using not only predictions of offensiveness but also predictions of rating discrepancy by gender within the annotators themselves. We show that both classifiers are effective at increasing rates of finding microaggression *and* that each finds different types. Finally, we show that self-reported and observed microaggression differ substantially in their types, suggesting computational mechanisms are needed to precisely estimate their true pervasiveness. Our work and corresponding data and models fill a key missing gap for expanding the range of offensive language detection to linguistically subtle comments.

Acknowledgments

The authors would like to thank the three reviewers for their very helpful and insightful feedback, Figure Eight for providing funding as a part of their AI for Everyone project, and the Washington Post data science team for data used during initial pilot studies of this project. YT gratefully acknowledges the support of the Okawa Foundation and of the NSF grant no. IIS1812327. EA was supported by NSF GRFP under grant no. DGE1745016. The authors would also like to acknowledge the people who helped with the pilot study: Gayatri Bhat, Alexander Coda, Anjalie Field, Shirley Anugrah Hayati, Arnav Kumar, Sachin Kumar, Shrimai Prabhumoye, Qinlan Shen, Sumeet Singh, Craig Stewart, Michael Yoder, Yiheng Zhou.

References

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Gavin de Becker. 1997. *The Gift of Fear*. Little, Brown.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3(Jan):993–1022.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Bradley Campbell and Jason Manning. 2014. Microaggression and moral cultures. *Comparative sociology*, 13(6):692–726.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM*, ICWSM '17, pages 512–515.

Maeve Duggan. 2017. Online harassment 2017. *Pew Research Center*.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *ICWSM*, pages 112–120.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #metoo stories. In *Proc. ICWSM*.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proc. ACL*.

Shandra Forrest-Bank and Jeffrey M Jenson. 2015. Differences in experiences of racial and ethnic microaggression among asian, latino/hispanic, black, and white young adults. *J. Soc. & Soc. Welfare*, 42:141.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Aparna Garimella and Rada Mihalcea. 2016. Zooming in on gender differences in social media. In *Proceedings of the COLING Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 1–10.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham, and Chris Callison-Burch. 2019. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*.

Samantha Jaroszewski, Danielle Lottridge, Oliver L Haimson, and Katie Quehl. 2018. Genderfluid or attack helicopter: Responsible HCI research practice with non-binary gender variation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 307:1–307:15. ACM.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16.

Youxuan Jiang, Catherine Finegan-Dollak, Jonathan K Kummerfeld, and Walter Lasecki. 2018. Effective crowdsourcing for a new type of summarization task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 628–633.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jackson Katz. 1995. Reconstructing masculinity in the locker room: The mentors in violence prevention project. *Harvard Educational Review*, 65(2):163–175.

Kiyun Kim. 2013. Racial Microaggressions. http://nortonism.tumblr.com.

Jia Li, Xuan Liu, and Min Sun. 2018. Research on gender differences in online health communities. *International Journal of Medical Informatics*.

Scott O. Lilienfeld. 2017. Microaggressions: Strong Claims, Inadequate Evidence. *Perspectives on Psychological Science*, 12(1):138–169.

Gabriel Magno and Ingmar Weber. 2014. International gender differences and gaps in online social networks. In *Proceedings of the International Conference on Social Informatics (SocInfo)*, pages 121–138. Springer.

Kevin L. Nadal, Katie E. Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling and Development*, 92(1):57–66.

Blake Neff. 2015. California trains individuals to avoid microaggressions.. http://dailycaller.com/2015/06/10/california-trains-professors-to-avoid-microaggressions/.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The masc word sense sentence corpus. In *Proceedings of LREC*.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics*, 4(1):61–74.

James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. Linguistic inquiry and word count: LIWC 2015 [computer software]. pennebaker conglomerates.

Lisa Posch, Arnim Bleier, Fabian Flöck, and Markus Strohmaier. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *arXiv preprint arXiv:1812.05948*.

Farzana Rashid and Eduardo Blanco. 2017. Dimensions of interpersonal relationships: Corpus and experiments. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2307–2316, Copenhagen, Denmark. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5, pages 311–321, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate:

developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

Anna Schmidt and Michael Wiegand. 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*, pages 1–10.

Kimber Shelton and Edward A Delgado-Romero. 2013. Sexual orientation microaggressions: The experience of lesbian, gay, bisexual, and queer clients in psychotherapy. *Psychology of Sexual Orientation and Gender Diversity*.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, and Xifeng Yan. 2016. On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Derald Wing Sue. 2010. *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. Wiley, Hoboken, NJ.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha M.B. B. Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, 62(4):271–286.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *LREC*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. pages 78–84.

Gloria Wong, Annie O Derthick, EJR David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and social problems*, 6(2):181–200.

Wei Xu, Alan Ritter, and Ralph Grishman. 2013. Gathering and generating paraphrases from twitter with application to normalization. In *Proceedings of the sixth workshop on building and using comparable corpora*, pages 121–128.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*, pages 2979–2989.