April 10

Towards a critical NLP: Fallouts of the algorithmic management of 'abusive' language (working title)

Maya Indira Ganesh and Zeerak Waseem

In December 2017, Seda Gürses, Francien Dechesne, Solon Barocas and XXX brought together a group of lawyers, computer scientists, communications scholars and cultural scientists to think through algorithmic bias and discrimination from an intersectional perspective. There has been relatively less attention paid to what intersectionality might mean as a theoretical, epistemological and computational approach to bias. (more about the workshop) This article emerged as a collaboration from two doctoral student-researchers – Maya Ganesh and Zeerak Waseem - present at that workshop and is drawn from conversations between them.

In May 2018, at a music festival in Alabama in the United States, a young woman was invited up on to stage by the Pulitzer prize-winning poet and musician, Kendrick Lamar, to rap along with him to *M.A.A.D City*. The track is replete with the N-word, and the woman -White – said it, so Lamar stopped her to say: "you gotta bleep one single word." At first she did not realise why she had been interrupted, and then she got it as the crowd booed furiously[1]. While some may argue that a word in rap lyrics can be said by 'anyone' because it is not directed as a slur at an individual, it is widely acknowledged that the N-word is only to be used by African American people because of the very unique context of the word's historical legacy and continued use in popular culture and everyday speech.

Meanwhile, governments around the world are increasingly implementing regulation for content moderation such as the Stop Enabling Sex Traffickers Act (SESTA) and the Fight Online Sex Trafficking Act (FOSTA) in the United States of America which seek to curb online sex trafficking, the German Network Enforcement Act (NetzDG) which aims to limit online hate speech, and the proposed European Union regulation on dissemination of terrorist content, which aims to curb the spreading of terrorist content online. All four of these acts assign liability to platforms for either hosting the offending content or failing to act to reports of it, which incentivises online platforms to apply computational means of detection to prevent being sanctioned.

However, applying computational approaches brings into question the sanity of the current state of computational methods and the bodies which they govern and rely on. Considering SESTA and FOSTA, advocacy groups argued that the bills would increase the risks of exploitation of sex workers and violence towards them (CITE HERE). NetzDG has been criticized for being "vague, overbroad, and turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal." (Human Rights Watch, 2018), and the proposed European Union regulation on dissemination of terrorist content has drawn critique from advocacy groups for the potential consequences it may have to (Kayyali, 2019).

In each of these criticisms, the people affected by the content moderation are at the centre of the critiques. NetzDG for stifling lawful speech with no ability to appeal for the users, the proposed regulation on dissemination of terrorist content for hindering the work in documenting human rights violations, and SESTA and FOSTA for endangering sex workers. To illustrate how issues on automated content moderation may influence and inadvertently discriminate, we consider the field of natural language processing (NLP) and hate speech detection. We argue that the elimination of bias in NLP systems may not be achieved by identifying and removing bias per se, but by introducing and reinforcing it.—

~~In this article, we argue that the elimination of bias in natural language processing (NLP) algorithms may not be achieved by identifying and removing bias per se, but by introducing it.~~ Taking the case of African American Vernacular English (AAVE) and drawing on the ongoing doctoral work of both authors, we show that the algorithmic identification and classification of

speech as ‘abusive’ on scale risks perpetuating biases against an already marginalised community. Therefore we call for a ‘critical NLP’ that takes an intersectional perspective on practices of algorithmic regulation of language.

The N-word, ending in -er, was used by White slave owners to refer to Black slaves; and Black slaves used it, with its schwa, that is without its /er/ ending but an /a/, to refer to themselves (Rahman 2012 p.138-9). But despite its negative historical reference and restrictions on its use, the word circulates freely in popular culture globally through Black/African American culture. Jacqueline Rahman finds that the word has unique cultural, historic and political meanings for African Americans: as counter-language, a form of solidarity, and as emblematic of cultural, affective and spiritual practices of survival under conditions of racism. (2012). She shows how African Americans have always used the N- word to refer to each other and their community, from the time of slavery to the present[2].For example, Rahman cites an old ‘slave holler’ by a woman called Harriet Jacobs who successfully evaded capture by White slave owners: “Dis niggers too cute for ‘em dis time” (p. number).

In the present day, Black comedy, Hip-hop and Rap use n*****a, b***h and h*e in the context of entertainment, art and culture to performatively present the conditions of racism, poverty, institutionalised violence and class discrimination that Black people in the US struggle through. Perhaps the most critical work that informs studies of African American culture and politics is the notion of ‘double consciousness’: that personal Black identity is shaped by communal Black solidarity as well as the national White supremacist ideology (du Bois cited in Brock, 2012 p 522). Thus it is not just about the N-word as a form of recognition within a community, but acknowledges the ongoing interpellation of African Americans as former slaves – evidence of the persistence of White supremacy in US society. In this context, then, while the N-word in Hip-hop may seem like ‘just entertainment’, it is in fact ‘ritual drama’ in the discursive construction of Blackness. Maybe some reference on language and ideology: https://www.annualreviews.org/doi/abs/10.1146/annurev.an.23.100194.000415

However, the N-word’s use in African American Verbal Expression (AAVE) and in popular culture presents an acute and particular problem for algorithmic speech and content management practices: algorithmic identification and monitoring of online speech cannot distinguish between a contextual use of this word (for example, when Black people might be speaking to each other, or rapping), and its use as a racial slur (for example, when someone might use N, B, or H words to be abusive to an individual or about a community). This is because working at scale, natural language processing (NLP) using machine learning (ML), cannot identify who is speaking and in what context without additional information beyond the text. As a result, AAVE gets classified as ‘abusive’ by NLP systems because of the presence of these words; this presents a case of further discrimination against an already marginalised community. Thus we claim that in order to identify abusive language, we have to introduce *contextual biases*.

Waseem’s research is focused on building ~~a~~ NLP classifiers to identify abusive language online. Waseem uses machine learning models as a tools to identify abusive language. These operate by observing patterns in labeled datasets, where an algorithm seeks to learn a function which allows for separating between the individual categories of data. For instance, if a dataset has been labelled for the categories “hate speech” and “Not hate speech”, a model will seek to learn the patterns which allows for distinguishing between the two classes.

In current and previous work, he has relied on combinations of communities of people both online (on crowd-work platforms like Figure8) and offline (for example, Danish feminist activists) to annotate tweets from large scale online events like Gamergate that generated significant abusive language and attacks both online and off (ref). Having annotated tweets for their abusive content – sexist, racist, and combinations of these – Waseem finds that it is an emotionally exhausting task even when none of it was directed at him personally. He says[3]:

"It really really got to me...the stuff that got to me were the calls for genocide, calls to 3-d print a gun and kill people of colour. It was tough. That made me aware of what professional moderators are facing but also people who faced things themselves, like the women at the centre of Gamergate - Anita Sarkeesian, Zoe quinn, Brianna Wu. I got to see what some of the harms were.

April 10

At some point, an ex-partner of mine, a White Swedish woman, was also annotating some tweets, and there were tweets about calls to kill Swedish women with men of colour."
Include: #yourslipisshowing - no one cared about this but when #gamergate became visible itwas because white women were affected by it.

Recent documentaries like The Moderators (Ciaran Cassidy and Adrian Chen, 2017)[4]and The Cleaners (Hans Block and Moritz Riesewieck, 2018)[5] have demonstrated the human, material, affective infrastructures of online content moderation and the toll they take on individuals. The case for computational approaches to online content moderation is a little like the case made for lethal autonomous weapons: that the psychological price for 'cleaning' the internet may be better absorbed by computational systems that do not 'feel', wear out, or need care and recovery as human soldiers do on battlefields. The Campaign to Ban Landmines argues that landmines were introduced precisely because they were considered 'autonomous' and did not require any human decision to activate, except the proximity of a human body to the mine's tripwire (in Suchman and Weber 2016 ref). Automated speech recognition systems can be a little like a landmine.

Consider these tweets[6]:

@twitterhandleredacted: it aint nothing to cut a bitch off
They Faggots @twitterhandleredacted: Orioles petty as fuck...got swept up outta this bitch
@twitterhandleredacted: That son of a bitch moment when your walking through your house in the dark and stub your little toe on the wall
@twitterhandleredacted: I can't stop being a lil bitch;
@twitterhandleredacted: When life knocks you down, stand the fuck up and say "You hit like a little bitch."

A casual survey of popular African American everyday speech and culture demonstrates that the word b***h may be used by women to refer to each other but not necessarily as an expletive nor as abuse. But an abusive language classifier that was trained on a data set that contained the word 'bitch', and if 'bitch' were annotated as an abusive word, then the vast majority of these tweets would be classified as abusive by its model. But not all of these are used as abusive slurs against individual women (however some of these are sexist).

When someone says 'that son of a bitch moment when your [sic] walking through your house in the dark and stub your little toe on the wall' they are expressing exasperation, not referring to any person as a bitch, or the son of one. 'You hit like a little bitch' however might be construed as 'sexist'; and 'ain't nothing to cut a bitch off' is a uniquely African American construction. It is for this reason that Waseem has asked people to volunteer to annotate tweets like these to clarify their context. It is also in this context that the other author of this article (Ganesh) started annotating some tweets in Waseem's database.

Z: I think this paragraph needs to be rewritten. It's not clear to me what you mean by "are to be classified" - the abusive label is already there, so we don't need to annotate them for that - but the end goal is to classify them as abusive or not and let that also depend on whether what they've written is AAVE or not - which is what we were annotating for.

Waseem has a dataset of 24000 tweets that are to be annotated as 'abusive' or not. These tweets were previously annotated by crowd-workers at the instruction of Thomas Davidson and colleagues (2017). It becomes rapidly clear that the tweets are replete with the N, B, H, P words, among others. Working through the dataset is to be immediately confused: for Ganesh, as an Indian raised on a media diet rich in US popular culture including Hollywood and Hip-hop, it was difficult to classify some tweets as 'abusive' (as AAVE or not) just because they contained the N-word or the B-word. Some of these were used in casual conversational tweets between known people, some of these are 'burns', other usage of the word was to shame, ridicule, or heckle. In most instances, people – assumedly African Americans - were referring to themselves, their friends, families, celebrities, neighbours, or just 'people', as n****s or b*****s. This is not 'abusive' speech, it is just how African Americans talk. More confusing to this Indian annotator is the bias of her cultural background, and feminism perhaps: the repeated occurrence of the word p***y in various contexts is confusing. Are these tweets misogynist, frivolous, or just humourous ? Are women talking to each other? Is this a broader social commentary about heterosexuality

April 10

heteronormativity? (include some P-word tweets to show, and for the reader to get a sense of them?)

Z: Just updating these numbers.
24K Annotated for hate, offensive, neither.
5009 annotated for AAVE:

2444 Tagged as AAVE:
2336 tagged as Offensive/Hate speech
108 tagged as not offensive/hate speech

2584 Tagged as Not AAVE:
1813 tagged as offensive/hate speech
750 tagged as not offensive/hate speech

Waseem finds that of 2444 tweets that he annotates as AAVE, only 108 are *not* labelled as offensive or hate speech by online crowdworkers, so the majority *are*. A classifier trained on this dataset annotated in this way will therefore start identifying legitimate AAVE as 'abusive'. The implications of this for an already-marginalised community are staggering: that how African Americans speak English on online platforms could effectively be deemed 'hateful' or 'abusive' by a ML system.

Z: First sentence should be changed here to reflect the changing of the title.
It is for this reason that Waseem pushes for critical NLP.
~~It is for this reason that Waseem claims that he wants to "kill his field", a dramatic and not-literal claim that underlies our argument.~~ That as an emphasis on the fairness of machine learning systems grows, there is a search for an accurate definition of what counts as, or introduces, bias or discrimination in such a system. However, he finds that in trying to sanitise the internet of something termed 'abusive' language, we risk taking words out of context without being able to acknowledge their highly specific, cultural resonance. And while introducing a human moderator to check onhow the classifier is working might be less 'efficient', we take this moment to ask why and how 'efficiency' has been constructed as valuable in the first place. (something more on the efficiency imperative). How do we integrate justice or equality as the requirements of a system? And these not to rendered in terms of rules set down by law, but by the contextual parameters established by history.

A general blindspot (better word?) in the NLP field, according to Waseem, is that the labelling of speech by crowd-workers would contain biases, however here we highlight that biases may be reinforced.
A general consensus in the NLP field, according to Waseem, is that the labelling of speech by crowdworkers would eliminate biases, however here we argue the opposite. A crowd-worker follows particular guidelines and protocols without necessarily knowing the context of how certain words are used and in what context (ref). Crowd-workers tasked to manage online content for Facebook for example, are given a set of guidelines that have been shown to be problematic at best, and disconnected from the reality of online violence at worst[7] (facebook files reference and footnote). The Guardian's investigation into the Facebook Files shows that their content moderation guidelines adhere closely to conventions established by US hate speech laws in which certain categories of people are protected to the exclusion of others; so sexual identity groups are protected but 'the poor' are not. And individual members of a religion or value system are protected but the ideas are not; so while violence against Muslims is not tolerated, 'anti-Islam' groups might be allowed. Incitements to violence against people based on their characteristics (being fat, thin, tall, red-haired, blonde etc) are not protected by these guidelines, but actual people are.

A critical NLP approach would acknowledge the value of contextual biases, that how an African American might annotate a dataset containing N, B, H or P words will be very different from how a crowd-worker online and elsewhere might label the same words. New research on content moderation strategies also suggests a direction towards what we propose. Caplan for example (2018 - ref) presents content moderation practices as 'community reliant', 'artisanal' and 'industrial' based on the mission, business model and size of content moderation

teams. They suggest that a nuanced approach that balances context-sensitivity with consistency is required to address online moderation practices going forward.

A critical NLP approach suggests that we have to be intersectional in addressing the problem of algorithmic discrimination. This might include justice considerations, and the historic context of discrimination against a group, to shape how a ML system is built and how the data is labelled. There may be practical and material considerations to this that platform companies may need to pay attention to. For example, despite information from activists and scholars about how Facebook was being used to promote violence against the Rohingya in Myanmar, the company did not increase the resources for Burmese language content moderation (Reuters investigation ref).

On Ganesh and Waseem as non-AAVE speakers doing and writing this.
Who we consulted and spoke to.
Waseem recruiting AAVE speakers to annotate datasets. Reiterate efficiency imperatives mentioned above.
But also that intersectionality operationalised as a kind of arithmetic is not the point- is precisely how diversity and decolonisation are words that now have little meaning. How this article / conference attempts to address intersectionality in a different way.

In his groundbreaking study of 'Black Twitter', Andre Brock (2012) challenges the normalisation of Whiteness in how we understand big data and social media, and in how we study of affordances of these systems for formation of community through speech. Social media is thought to be characterised by 'context collapse' (Marwick and boyd 2011- she intentionally does not capitalize the first letters of her names) in which all identities and relationships are flattened into uniform, 'univocal' ones; Brock argues however that the 140 character limit of Twitter actually encourages Black users to construct a discursive racial identity. He argues that "Black discursive styles and cultural iconography" are constructed through "cultural touch points of humor, spectacle, or crisis" (357) in a way that "subverts mainstream expectations of Twitter demographics, discourse and utility." Brock's work pushes us to consider how the introduction of AAVE speakers in online annotation processes might offer an entirely different perspective on n- and b- words. Waseem's work on this project now looks in exactly this direction as he recruits AAVE speakers as annotators of the dataset containing AAVE. In this way, critical NLP would necessarily find ways to include historically marginalised groups of people to be present and engaged as experts on their own culture and informing the development of ML systems. [some nice ending sentence]

Zeerak needs to add the piece about what it means to get AAVE speakers now -

Bios.

1https://www.bbc.com/news/newsbeat-44209141

2It is also true that many African Americans eschew the use of the N-word, including well-known cultural icons from Ice Cube to Richard Pryor. Rahman notes that a particular generation of African Americans who found community and identity through the church (for example) and grew up around Civil Rights movements in the 1960s may not appreciate its use. A younger generation, she says, has perhaps 'reclaimed' the word in a show of defiance, as well as to critique the role of traditional African American social, cultural and community institutions.

3Interview with Maya Ganesh, October 8, 2018.

4https://vimeo.com/213152344

5http://www.gebrueder-beetz.de/en/productions/the-cleaners#synopsis

April 10

6Tweets are all from – Database details.

7https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules

Human Rights Watch, Germany: Flawed Social Media Law: NetzDG is Wrong Response to Online Abuse, February 14, 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Kayyali, 2019, WITNESS brings together voices to push back on dangerous EU “Dissemination of Terrorist Content” proposal, Dya Kayyali