Insights //

alt.advisory

Context matters: The importance of local nuance in online content moderation

By Phenyo Sekati*

17 July 2023

ABSTRACT

This article explores the challenges of defining hate speech solely based on content without regard to context, which leads to inconsistent and inequitable regulation. It examines the importance of considering local political realities and power asymmetries between social groups, as well as the impact of cultural and linguistic nuances on content moderation. Lastly, it offers recommendations for improving the process and ensuring fair and effective regulation.

Citation: P Sekati, *Context matters: The importance of local nuance in online content moderation*, ALT Advisory Insights 2023 (3) (17 July 2023).

+-*#*-+

INTRODUCTION

Context and nuance play an important role in determining the bounds of hate speech. Consider the example of "Kiss the Boer, Kiss the Farmer", a political chant popularised by the Economic Freedom Fighters (EFF), a South African political party. Afriforum, an Afrikaans civil rights organisation, sought to have the song declared as hate speech, contending that the chant is a violent derivate of "Kill the Boer, Kill the Farmer" (*Dubul' ibhunu*) an Apartheid-era struggle song, and argued that it encouraged the killing of white farmers. The EFF maintained that before democracy, the song was directed at the dispossession of land during Apartheid, and is currently directed at ongoing land injustice. Moreover the EFF argued that political songs should not be interpreted literally. Determining whether the song amounts to hate speech accordingly requires a nuanced understanding of the historical, political and racial context within which it was sung. Our courts grapple with this context and apply an objective standard when they make such a determination.

But what happens if someone posts "Kiss the Boer, Kiss the Farmer" online? Out of context, the words alone may seem harmless. But will those tasked with moderating content on social media consider the historical, political and racial context when assessing whether the post violates its community standards? Are social media companies equipped to understand the necessity of contextual comprehension, particularly in diverse cultural and linguistic environments like South Africa?

This article explores the significance of nuance in online content moderation and looks at some of the challenges of regulating hate speech.

UNDERSTANDING THE REGULATION OF HATE SPEECH IN SOUTH AFRICA

In South Africa, everyone has the right to freedom of expression. But this doesn't mean you can say whatever you

want – the right has limitations. For example, section 16(2) of the Constitution excludes the advocacy of hatred based on race, ethnicity, gender, or religion, that incites harm.

There is a further prohibition of hate speech in the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000 ("Equality Act") and there may soon be another if Parliament passes the Prevention and Combatting of Hate Crimes and Hate Speech Bill ("Hate Speech Bill"). These laws define hate speech as intentional communications that could be reasonably construed to demonstrate a clear intention to be harmful or to incite harm, *and* to promote or propagate hatred on a broader list of grounds including race, gender, colour, culture, and sexual orientation. The use of the word "intention" raises practical difficulties, and both the Equality Act and the Hate Speech Bill create exemptions for communication that falls, for example, under *bona fide* artistic expression, scientific inquiry, or in fair and accurate reporting in the public interest.

// "The stakes are high, and the opportunities to get it wrong are endless."

As evidenced by the "Kiss the Boer" chant, these may not be simple determinations. However, our courts are able to make such determinations by considering the social and political contexts wherein language is used, and by relying on precedent to understand the bounds of freedom of expression.

Courts are equipped to make these determinations and do so in a way that strikes an appropriate balance between freedom of expression and competing rights such as privacy, dignity and access to information. But courts are not the only forum where such determinations are considered. There has been a significant shift in the dissemination of hate speech from real-world interactions to social media platforms. This shift has transferred the responsibility of identifying and addressing hate speech from the courts to content moderators employed by multinational social media companies. So how do they do it?

HOW SOCIAL MEDIA PLATFORMS REGULATE HATE SPEECH

Most social media platforms have developed 'terms of service' or 'community standards' that regulate the types of content that are permitted or prohibited on their platforms, which typically include a prohibition against hate speech. TikTok, for example, does not allow "any hateful behaviour, hate speech, or promotion of hateful ideologies." Users are required to accept the terms to participate on a platform, and there are consequences for non-compliance. A user who posts prohibited content could have their post removed, or buried (so it is seen by fewer people), or have their account suspended or banned.

When these companies get it right, they protect users from harm and harassment, but when they get it wrong, they may violate a user's rights to freedom of expression and access to information by being overly restrictive – or fail to protect other users' rights by failing to act on legitimately harmful content. The stakes are high, and the opportunities to get it wrong are endless. In the last quarter of 2022, Facebook reportedly removed 11 million pieces of content that contained hate speech – and that number doesn't include content which violated a different provision such as violence or misinformation.

Because of this proliferation of content, social media companies rely on automated tools to flag and remove content. According to Facebook – "[their] technology finds more than 90% of the content that [they] remove before anyone reports it." While automatic filters and machine learning models have enabled platforms to process large amounts of content, they bring additional challenges, including their inability to address nuance.

BIASES IN AUTOMATED CONTENT MODERATION

Major social media platforms rely heavily on automated tools such as Natural Language Processing (NLP) to detect and

remove harmful text-based content. However, embedded human biases in the technology can lead to inadequacies when moderating content in a country as linguistically, culturally, and ideologically diverse as South Africa.

Of the five billion people using the internet today, 75% are from the Global South, which is home to 90% of the world's 7000+ languages. Many of these marginalised languages are, however, not supported on most operating systems including language processing software. The lack of inclusive language processing software similarly applies to the use of slang in many communities as the context of the text and the area in which the post is published may not be considered.

Even in widely spoken languages like English, NLP tools struggle to interpret the surrounding context or adapt to novel cases thus limiting or removing speech that may be constitutionally protected, or failing to identify speech which may be harmful. Moreover, NLP tools are heavily dependent on their training data and any biases it incorporates. For instance, in 2018, YouTube's algorithms erroneously classified a discussion about black and white chess pieces as being "harmful and dangerous". Another example can be seen with the banned use of the words "lesbian" and "gay" on TikTok as well as the censorship and shadow banning of LGBTQ+ hashtags and content.

The issue of local nuance is exacerbated in South Africa where the use of language and symbols have different meanings to different cultural and ethnic groups, with hate speech often being determined by the *context* of the communication rather than the *content*. For example, the word "boer" (which translates to "farmer") can be interpreted to be both offensive and complimentary by Afrikaners in South Africa depending on the context. Words such as "bobbejaan" (which directly translates to "baboon" in Afrikaans but has been used to deride black people) and "meid" (which directly translates to "maid" but has been used to demean black or mixed-raced women) could easily be subject to either under- and over-removal which, as a result, would either suppress the rights to freedom of expression and access to information or perpetuate the spread of hate speech online.

These examples highlight the importance of cultural context and the need for content moderation policies and tools that are sensitive to the specific historical and cultural nuances of different communities. The political realities, power asymmetries between social groups, and cultural and linguistic nuances on content moderation must always be considered.

// "The issue of local nuance is exacerbated where hate speech is determined by the *context* of the communication rather than the *content*."

For example, the decision of South African courts to ban the public display of the apartheid flag of South Africa was rooted in its symbolic ties to a system of racial segregation and oppression, which far-right groups have used to promote white supremacist ideologies. However, if social media platforms do not understand the historical background of the apartheid flag in South Africa, they may not realise the harm caused by its use and may fail to take appropriate action to remove any publications of the flag.

In cases where hate speech is excessively policed through automated content moderation, the rights of marginalised groups and languages are disproportionately compromised. This results in infringements upon their rights to equality, non-discrimination, and freedom of expression. Where these filters fail to detect nuanced forms of hate speech, such as with the use of the apartheid flag or specific derogatory terms like "bobbejaan" and "meid" when

referring to certain individuals, marginalised groups are more susceptible to discrimination and exclusion.

CONCLUSION

Content moderation is a complex task, and the bounds of hate speech can be particularly difficult to define and moderate without taking into account its context. As language and cultural nuances evolve, it is essential to have dynamic responses to hate speech, in collaboration with experts with contextual knowledge of speech, in order to strike a balance between moderating hate speech and protecting freedom of expression.

According to the UN Human Rights Committee's General Comment 34 on Freedom of Opinion and Expression, social media platforms must ensure that restrictions are applied only "for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicated." To effectively make these determinations, social media platforms should employ a pluralistic model of content moderation starting with the provision of sub-regional or country-specific lists of hateful expressions.

Local moderators who speak the local dialect of the language and who are culturally and linguistically close enough to the sources of the posts should be appointed in every region to assist automated tools in moderating content. Lastly, user participation is also crucial to ensuring that content is not disproportionately moderated and social media platforms should ensure that users have effective opportunities to appeal against decisions they consider to be unfair.

+-*#*-+

* *Phenyo is a Tech Rights fellow at ALT Advisory working the intersections of human rights and digital technologies in the African region.*

ENDS.