Hate speech is typically categorized in one the following ways: non-abusive, abusive, argumentative, counter-speech, causal profanity, hate speech, offensive language, explicit abuse, or implicit abuse when annotating a social media post. These types can be categorized as multiple of these or say that they need more information. These methods of categorization, while not perfect when done by machine-learning systems, can be somewhat effective; the problems arise when more elements, such as emojis or images, come into play. The complications that arise in detection when things other than words are included are fundamental to understanding when it comes to analyzing hate speech online.

Emojis, images, and other modifications of words can disguise hate speech in a way that successfully evades machine-learning systems detection of hate speech. In addition, certain contexts may change how hate speech shold be interpreted, transforming it it into something like offensive language or counter-sppech. For example, the usage of the monkey emoji when used in relation to Black athletes in sports is a form of implicit hate, but is not able to be detected as such by the machine because it does not understand the underlying meaning of the emoji.

The *Reclaiming Arabic Words* case showcases this difficulty. The clurs there were being used by the account, while offensive, were not hate speech. The account was being used for "discussing queer narratives in Arabic culture" and that social media has been a place for the LGBTQ+ community to express themselves. The post was not meant to be hateful, but instead empowering. The post while it was using slur terms, the content was utilized in the way of counterspeech making it an acceptable post. The machine's misclassified the post as hate speech when it was really counterspeech becuacse it failed to interpret the social and cultural context of the post. These two examples demonstrate how emojis and images can disguise or make certain

speech alright, and effectively evade machine-learning systems detection or categorization of such speech.

Pivoting to the question of the human rights reposonsibilites of social media companies undoubtedly social media companies have human rights responsibilities to identify and respond to hate speech. Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR) states that “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence” is prohibited. Therefore speech that incites this advocacy is prohibited and social media companies are empowered and obligated to identify and respond to such hate speech under the ICCPR. Additionally, Article 10 of the European Human Rights Doctrine (ECHR) states that free expression allows a limit “for the protection of the reputation or rights of others.” This doctorine it states that there is a right to free expression but not at the expense of others rights. Social media companies are therefore justified to take action when it comes to limiting free speech, and removing content, when that speech comes at the expense of others rights, and recognizes that individuals have a right to be free from hate speech.

The changing nature of hate speech makes it increasily difficult to be detected by machine-learning systems demands a response by companies. Current annotation methods fall short when analyzing content beyond words, such as emojis or images. Companies are obligated, under the ICCPR and the ECHR, to protect the individuals who use their platforms and are given the power to remove content if they are violating the rights of others, and are expected to use such power to protect the human rights of indiviudals to be protected from hate speech who use their platforms.