Public Reaction to Controversial Art: A Study on Lil Nas X's Montero Era

Anonymous Submission

Abstract

With the prominent shift that social media platforms have taken over the course of the past years from a means of communication and connection to a battlefield for public debate, there has been witnessed a significant spread of hate speech. This study provides a quantitative analysis on the public's reaction to art that is controversial in aim of understanding how open different communities are to such content. We perform sentiment analysis on tweets from over 650,000 users collected in the time span during which Lil Nas X, a black gay artist, started his Montero album era during 2021. An XL-Net classification model is then applied to the same tweets to study the public reaction to his controversial art and analyze the degree of hate that these users portrayed. We finally discuss the dispersion of these hate tweets across different regions of the world and analyze the different reactions from metrics of homophobia, racism, and extreme conservatism.

Introduction

Social media platforms such as Tumblr, Facebook, Twitter, and Reddit have always been an open stage for the sharing of public opinion and perspective ever since their inception. With the great number of users on each of these platforms, and despite tremendous efforts by their developers to always preserve a safe space for all users, there has been a lot of critical focus on the increase of hate speech and bullying that goes hand in hand with the expansion of said platforms, and the difficulty of its regulation. In the United States alone as of October 2021, there are more than 77.75 million active users on Twitter alone (Statista 2021). Given that, Twitter has been a major platform for political and ideological debate across the world and serves as an insightful hub for data that is representative of the public opinion.

According to the U.S. Council on Foreign Relations, the United States – although legalized same-sex marriage and advanced some rights for members of the LGBTQ+ community – today still faces a big challenge when it comes to equality (Angelo and Bocci 2021). Furthermore, there are nations in which homosexual relationships are still criminalized legally. Since the election of President Donald J. Trump in 2016, LGBTQ+ rights have been deteriorating and the hate against members of this community has become more widespread. This led to more people voicing out hate against members of the LGBTQ+ community.

Because of the number of people who follow up with public figures, Litchfield et al. (2018) examined the fan-athlete relationship in extensive research after the public hate that star athlete Serena Williams faced during the 2015 Wimbledon Championship. Using data collected from both Twitter and Facebook, they found a significant amount of hateful negative posts about Serena who was ridiculed for her masculine appearance and bullied for being a black player. Therefore, the hate speech directed towards Serena was in fact underlying racism and sexism. This was motivation for this study to approach a similar topic from a different angle, which will study the public reaction in regard to a new artist's work that sparked public debate not only from a race point of view, but also a religion and sexuality one. Instead of analysing homophobia and racism in a broad sense, we will be studying them as a reaction to specific events centered around the timeline during which American gay black artist Lil Nas X started advertisement for and released his debut album "MONTERO." Because all events related to this album release (music videos, collaborations with major retailers, live performances, interviews) have attracted a lot of debate on social media given their controversial content, this offers a very insightful data set that could allow the rather more accurate analysis of racism and homophobia around the world today.

Given the background and motivation behind this research, the problem statement is formulated as such: Was the public's reaction to the content released by gay black artist Lil Nas X indicative of a currently prevailing attitude of racism and homophobia, and what does that reflect about the marginalization of minority groups across the international community?

Related Work

The problem of online hate speech has been discussed at length for as long as the internet has been widely available to the public. Mehrabi et al. (2021) provided an excellent starting point for the previous work that has been done within the realm of applying natural language processing techniques to detect hate speech. Silva et al. (2016) categorized the targets of general twitter hate speech. Out of 512 million tweets posted between June 2014 to June 2015, they were able to identify 20,305 tweets. Of these tweets, they found 48.73% of hate-based speech could be categorized as race related,

followed by 37% for behavior-based reasons and 3.38%, 1.86%, and 1.08% for physical, sexual orientation, and class respectively. Other categories such as ethnicity, gender, disability, and religion were all below 1%.

XLNet (Yang et al. 2020) provides an improvement to BERT based models by combining the benefits of autoregressive language modeling and autoencoding techniques. This technique follows the traditional path of pretraining on an extensive unlabelled corpus but instead of following the route of autoregressive techniques and factorizing the likelihood of forward or backward products, it instead maximizes the likelihood of all permutations in a sequence allowing full contextualization of each token. This and other improvements lead to model training that outperforms previous methods such as BERT by a considerable extent. Subsequent research has found XLNet's success in other areas have translated well to classifying sensitive user specific data (Zhang et al. 2021), as well as generalized detection of hate speech from large corpora of twitter data (Mutanga, Naicker, and Olugbara 2020).

Method

With the aim of understanding the general population's perception regarding the art work produced by Lil Nas X, a collection of different approaches were utilized.

Data Collection

We collected tweets that strictly mention and/or discuss the artist Lil Nas from specific periods of time during which Lil Nas was receiving reactions from the public regarding his work. The data collection program utilized the Academic Access to Twitter API V2 which provided access to historic Twitter data. Upon collection of tweet objects, a comprehensive database was developed including information regarding tweet text, location, language, dates, and others. Using the GeoPy libary, location were retrieved from user profiles.

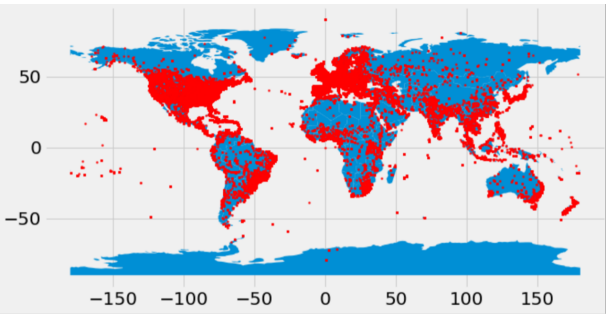

Figure 1: Geographic Distribution of Tweets Collected.

Sentiment Analysis using VADER

Vader (Valence Aware Dictionary for Sentiment Reasoning) was chosen in this study for several reasons. It is sensitive to both polarity (negative/positive) and intensity (strength) of the emotion underlying the text (Beri 2020). Furthermore, this model accounts for sentiment of statements and not only words, meaning that a positive word that is used in a negative context (such as the word "happy" in the context of the statement "I have never been happy") will not lead to a positive label on this statement. In addition to that, VADER also accounts for the meaning behind capitalization, punctuation, and repetition, which makes it the more appropriate model to use when analyzing tweets.

Hate Speech Classification using XLNet

For labeled twitter hate speech, we used the Davidson et al. 2017 data set. This consists of 24,000 tweets labeled as either hate speech, offensive or neither. These tweets consist of the English language and posted within 2017. Based on previous research on speech analysis and labelling (Zhang et al. 2021), we chose to build our model using the XLNet framework as it showed the best results in topic classification, and specifically hate speech (Mutanga, Naicker, and Olugbara 2020). We then designed our model to label a tweet as either hate speech or not hate speech. This was done with the XLNet-Base System which was chosen based on hardware constraints but also found to return acceptable results. Fine-tuning was completed with the Davidson et al. data set. The performance of the resulting model is shown in Table 1.

Table 1: Class Specifics.

	Precision	Recall	F-1 Score	Support
Hate	0.96	0.90	0.93	542
Not Hate	0.99	0.99	0.99	6938

Experiments & Results

The following sub-sections will present the results.

Hate Speech Classification

After running our model over the 650,000 tweets, we randomly selected 600 records and manually classified them resulting in Table 2. This result showed an overall accuracy of 93.79% +/- 4% at a 95% confidence level. The drop in metrics is not surprising given that the model was trained off an entirely different source of tweets as compared to the Lil Nas Dataset. Both share many commonalities. However, the tweets surrounding a specific pop star from a different set of years is going to result in language being used in ways that the model had not been exposed to during the training process. Further, a comparison of the output revealed an additional layer of complexity surrounding hateful terms. The environment in which we are identifying hate speech is one that discussed LGBTQ+ and race related topics extensively. The large rate of discussion surrounding these topics and the gap in generational linguistics and slang of those discussing these topics increases the number of tweets that are difficult to discern true intent.

Upon review, some errors were from the discussion of hateful terms being used against a community where a twitter user showed shock or resentment of their use. This process, where a word that historically would be classified as a slur, is instead re-appropriated by the targeted community to take its power away from those oppressing them (Coles 2016). This process creates very challenging situations for

natural language processing techniques where the acceptance of its use is in an ongoing debate. These results - while not ideal – are acceptable and provide insight. The model emphasizes the importance of the recall metric in this situation. Misclassifying hate speech as innocuous is dangerous to marginalized communities and our intention was to ensure that as much as possible could be identified. After this adjustment, the classification statistics were revisited to produce the results in Table 2.

Table 2: Revisited Class Specifics.

	Precision	Recall	F-1 Score	Support
Hate	0.49	0.80	0.61	37
Not Hate	0.99	0.95	0.94	562

Analyzing the Intersection of Classification and Sentiment Analysis

We decided to analyze the compounded results of XLNet classification and VADER Sentiment Analyzer as a new approach to drawing any trends in public reaction. It is obvious that Lil Nas X triggered hateful conversation through his album. However, the sentiment of this hate is under study. The negative reaction is categorized into two different aspect one that is hateful against the actual work presented by Lil Nas X, and the other is that reacting to this hate against him. Therefore, we can conclude from this that the public reacts negatively, and with increased hate, to controversial art irrespective of whether or not they're supportive.

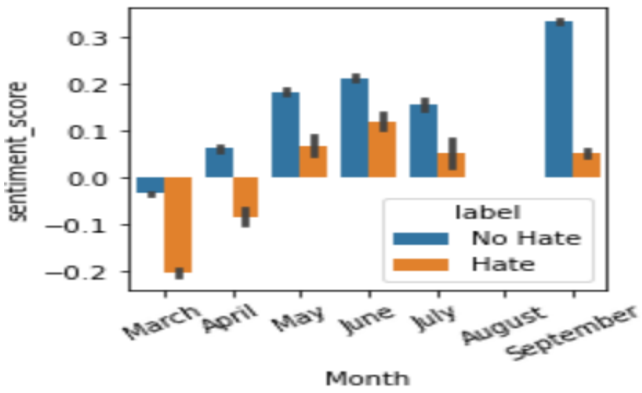

Figure 2: Compound Sentiment Scores with Hate Labels from a period of 2021.

An interesting observation from Figure 2 is that we notice hate tweets with positive compound sentiment in May and June. This is an indicator that even though the tweets do hold hate, they are not necessarily of negative sentiment - meaning that users were possibly defending Lil Nas X's content. In August, tweets were were part of the general conversation about art that Lil Nas X was releasing and we observe discussions with no hate attribution.

Hate Speech Categorization

Tweets labeled as hate speech were then categorized into several groups so that the change in topics can be viewed as events progressed through the year. Keyword classification was used to group tweets into the topics of race, sexual orientation, disability, religion, violence or other. The daily proportions of each category were plotted to show how the changes in discussion topics changed throughout the year.

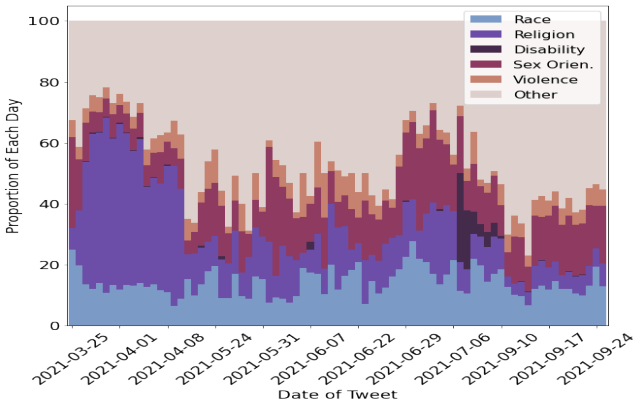

Figure 3: Daily Change in Type of Hate Speech.

Based on Figure 3, we see in late March to early April the dominance of the religious category which holds firm for several weeks as the activity surrounding Lil Nas's "Montero (Call Me by Your Name)" and the "Satan Shoes" garners the bulk of the attention. However, once that story plays its course, the discussion drops off very quickly. The next major change in focus comes in late June corresponding to the BET awards and the release of Lil Nas's "Industry Baby". Both events are areas where Lil Nas was very public about his sexuality and portions of the public reacted in hateful ways directed towards that self-expression. Through the end of this research's scope period (October – November), the proportions of hate speech remained largely consistent. This occurs through the release of his album "Montero" and single "That's what I want" on September 17th. In contrast to previous releases, we don't see the same spike in hate-related speech of any category. There is a spike in the volume of hate related speech corresponding to the release of the album. However, no category of hate dominates the conversation as others had previously Another interesting insight is the spike in race related hate, followed by sexual orientation at the end of October 2021. These events correspond to (1) Atlanta naming October 20th "Lil Nas X" day and then (2) rapper Boosie Badazz posting the previously mentioned homophobic tweets. While unconnected to each other, the timing shows how positive and negative events surrounding his work often end up drawing the attention of people who wish to say hateful things either way.

FP-Growth & Word Clouds

Using the tweets labeled as hate, frequent pattern mining using FPGrowth (Han, Pei, and Yin 2000) shows a mix of the various topics, communities, and events surrounding the scope of this analysis. Included are different communities such as "Christians" and "gay people". There's extensive use of the term "kids" referring to some people's expectations that the artist held up a family friendly image. "Boosie"

Figure 4: Word Cloud of Hateful Tweets.

refers to the rapper "Boosie Badazz" who had an ongoing disagreement with Lil Nas and authored several homophobic tweets directed toward Lil Nas. As well as many instances of self-identification ("I", "I'm") which show the intrinsic self-inserting nature that social media is based on. Most notably is a volume of terms related to the devil, Satan, hell, and shoes which connect to the controversial music video for the single "Montero (Call me by Your Name)" which involves many homoerotic actions between the artist and a character understood to be Satan. This video, released on March 26th was followed up with the release of what was quickly called the "Nike Satan Shoes". A series of 666 Nike sneakers which had been altered with satanic imagery and reportedly a single drop of human blood. This release by Lil Nas's production company had no backing by Nike itself, but these results show that the company and these shoes became very closely connected within the public's eye.

In order to gain further understanding of the frequent patterns detected, word clouds of the all tweets labeled as hateful were generated.

The word cloud in Figure 4 presents many words that were the center of discussion around Lil Nas X's Montero era. This confirms that the public reaction was mostly focused on Lil Nas X's sexuality, beliefs, and race more than they were concerned about the content of his songs. The interesting outtake from this is the fact that homophobia was one of the most used words overall, in addition to gay. Furthermore, the discussion of Christianity, the devil, and hell was also at the heart of the discussion. In this light, children were brought into the conversation showing that there was a huge spectrum of opinions on how children would be influenced by content such as this, to which the artist responded in a tweet of his stating that he is not responsible for raising people's children. The term "nigga" was used very frequently, showing that the use of this word on Twitter, despite their team's effort to stop hate speech, is still very prevalent.

Location-Based Analysis

We also studied the difference in perspectives from different countries as a reaction to the Lil Nas X Montero Era. The top five countries/regions with hateful tweets were (in descending order): United States, United Kingdom, Canada, France, and Taiwan. In order to understand whether these countries discuss the Lil Nas material form different perspectives, a word cloud was generated for each of the tweets from each of these countries, one with tweets labeled as non-hateful and one with tweets labeled as hateful. Upon careful analysis of all the word clouds generated and cross-comparing them with the sentiment analysis results, the below observations were made:

- US users would not want children being exposed to content such as Lil Nas X's which shows that there is an active attempt to raising more conservative generations unlike what the media portrays today. This is similar to the case of Taiwan which is a rather traditional/conservative.
- French and Taiwanese users criticize Lil Nas for his "islamophobia" rather than merely his sexuality and/or race. This presents an interesting shift of conversation when compared to North America (US and Canada) and shows the focus of different communities.

Conclusion

After analyzing tweets of over 650,000 users from all regions of the world to analyze their reaction to the controversial work of gay black artists Lil Nas X, it could be concluded from this study that despite the different cultures, backgrounds, and geographies involved, the international community is yet to overcome the boundaries of race and sexual orientation. Over the timeline of study, the data analysis on our classified tweets showed that hate indeed increases with the release of new controversial art.

An interesting outlook on this is that the reaction in North America, and the U.S. more so than Canada, was heavily infested with religious conservatism, homophobic slurs, and racial offenses. Interestingly enough, France and Taiwan, although demographically, racially, and culturally different, showed that they judged this artist based on his ideological history more than they did his race or sexual orientation (this does not eliminate that the latter also occurred).

In summary, analyzing the reaction that society has to works of art produced by a person of color who is also a member of the LGBTQ+ community provides a lot of insight into the actual state of progress that this society lives in today. Despite all the openness and acts of progressiveness portrayed in today's media, minority groups are still at threat across the globe. Whether in the United States or in Taiwan, the hateful reaction to a famously gay black man shows that society is yet to overcome the differences that are built-in within it. Although this analysis is focused on hate based on Gender and Sexuality minorities and racial groups, this shows the status of minority groups around the world and the hate threat that they live through on a daily basis.

References

Angelo, P. J.; and Bocci, D. 2021. The changing landscape of Global LGBTQ+ rights.

Beri, A. 2020. Sentiment Analysis Using Vader.

Coles, G. 2016. The Exorcism of Language: Reclaimed Derogatory Terms and Their Limits. *College English* 78(5): 424–446. ISSN 00100994.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining Frequent Patterns without Candidate Generation 29(2): 1–12. ISSN 0163-5808.

Litchfield, C.; Kavanagh, E.; Osborne, J.; and Jones, I. 2018. Social media and the politics of gender, race and identity: The case of Serena Williams. *European Journal for Sport Society* 15(2): 154–170.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54(6): 1–35.

Mutanga, R.; Naicker, N.; and Olugbara, O. 2020. Hate Speech Detection in Twitter using Transformer Methods. *International Journal of Advanced Computer Science and Applications* 11(9).

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the Targets of Hate in Online Social Media.

Statista. 2021. Twitter: Most users by country.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding.

Zhang, Y.; Lyu, H.; Liu, Y.; Zhang, X.; Wang, Y.; and Luo, J. 2021. Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study. *MIR Infodemiology* 1(1).