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

Mohamad Ali Kalassina

Matthew Johnson

Georgen Institute for Data Science

University of Rochester

Rochester, New York

DSCC440 – Data Mining

Fall 2021

Submitted: December 11, 2021

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 research project aims to provide a quantitative analysis on the public's reaction to art that is controversial. The study entails performing sentiment analysis on Twitter data collected from the timespan during which Lil Nas X, a black gay artist, started his Montero album era, which spanned over the course of several months in 2021 and included performances, merchandise, and the release of several videos. XLNet, an advanced classification model is applied to twitter users to study the public reaction to his "controversial" art and analyse the degree of hate that these users portrayed. The dispersion of these hate tweets is also depicted across different regions of the world to observe the different reactions from metrics of homophobia, racism, and extreme conservatism.

Table of Contents

List of Tables

List of Figures

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

I. Introduction

A. Background & Motivation

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. As observed in figure 1 below, in the United States specifically 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.

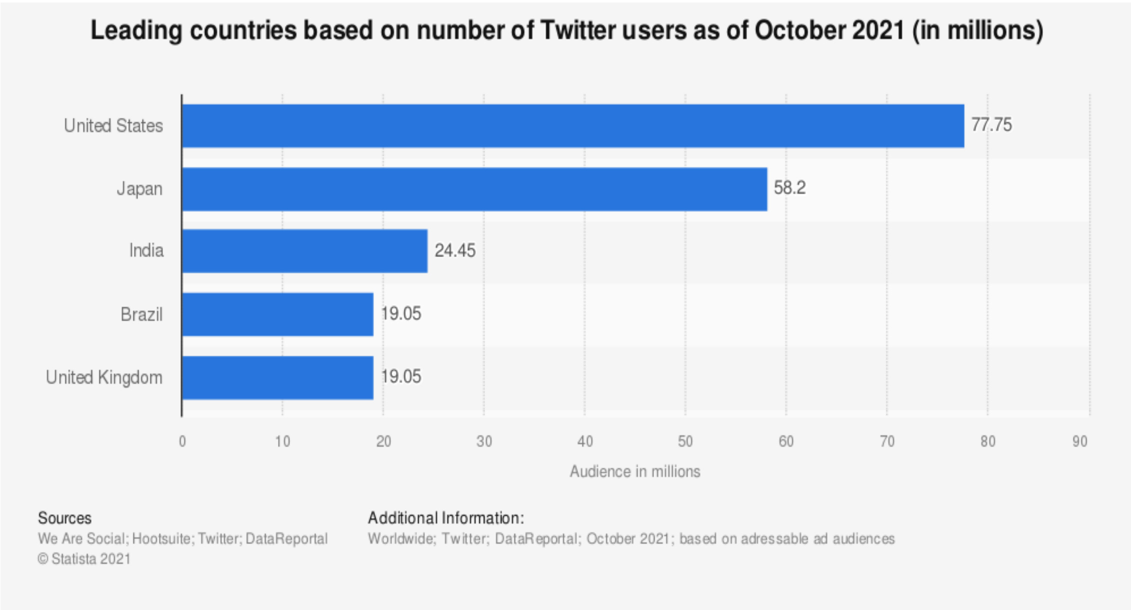

Figure 1: Twitter Audience per Country (in millions)

Many people live under the assumption that with 2022 almost here and because some legislation and public opinion around the world has become more tolerant and accepting of the LGBTQ+ community and people of color, that this is the general widespread norm. However, there are ample facts that prove otherwise. 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 genuine equality (Angelo & Bocci, 2021). Furthermore, there are nations in which homosexual relationships are still criminalized legally and frowned upon societally.

Since the election of President Donald J. Trump in 2016, LGBTQ+ rights have been deteriorating and the normalization of hate against members of this community has become more widespread, especially after some of his personal tweets denouncing the "excessive" rights that queer people have in the U.S. This led to more people voicing out hate speech against members of the LGBTQ+ community, including public officials who were appointed ever since. Therefore, this serves as a major motive to analyse the widespread of homophobia through Twitter.

Because of the number of people who follow up with public figures, be it artists, politicians, or others, certain actions made by these figures are often faced with a wide range of reaction from the public. 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. The findings of their analysis indicated that Serena 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.

Given that there exist numerous studies about hate speech and bullying on social media, this research intends on offering more accuracy by utilizing the improved Twitter API accesses to historic data. Therefore, instead of analysing homophobia and racism in a broad sense, we will be studying them as a reaction to specific events. These events will be centred around the timeline during which American gay black artist Lil Nas X – also known as Montero – started advertisement for and released his debut album "MONTERO." From here on by throughout the study, artist Lil Nas Z will be referred to as Lil Nas, the album name will be referred to as Montero, and the timeline from which he started advertising for this album until he finally released it will be referred to as the Montero Era. 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.

B. Problem Statement and Significance

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?***

The importance of addressing this problem lies on many levels, the most important of which is understanding the general public's tolerance and acceptance in the 21st century. Given the large diversity that the world holds today and the increase in marginalization of so many groups of different people, identifying a current surge of racism and homophobia is useful for the public sector and the private sector to take initiatives on making all spaces more inclusive, in effort to change the perception of those who still showcase hateful behaviour regardless of whether they are a majority or a minority. Furthermore, this could provide a motive for increased protection of marginalized groups of people, even those who are not a center of focus in this study.

II. Related Work

The problem of online hate speech has been discussed at length in both public and academic circles for as long as the internet has been widely available to the public. Mehrabi et al (2021) provides 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. Previous work (Silva 2016) focused on taking a general review of twitter hate speech and categorized the targets of that 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 ranged were all below 1%.

XLNet, proposed by Yang et al (2020), provides an improvement to Bert based models by combining the benefits of autoregressive language modelling 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. In addition, contrasting previous autoregressive techniques XLNet does not use a masking element in the pretraining process which previously caused discrepancies in the finetuning stages. These and other improvements lead to model training which 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 consumer specific data. (Zhang 2021) As well as to generalized detection of hate speech from large corpora of twitter data. (Mutanga T. 2020)

Earlier we discussed the work by Litchfield et al. (2018) on the discussion of Serena Williams at the 2015 Wimbledon Championship. A similar study of event-based hate speech research was completed by Williams and Burnap (2017) on the online reaction to the 2013 Woolwich, London terrorist attack. Both papers showcased the way authors of hate speech can feel emboldened by a current event and target their hate at a specific person or marginalized community.

III. <u>Data Collection & Analysis</u>

In this section, a detailed description of the below is presented:

- Data collection process
- Data pre-processing
- Exploratory Data Analysis

A. *Data Collection Process*

The data collection process entailed collecting tweets that strictly mention and/or discuss artists Lil Nas. Furthermore, the tweets were retrieved from specific periods of time during which Lil Nas was receiving reactions from the public regarding his work. The data collection program created (*Data Collection - LilNasX.py*) utilized the Academic Access to Twitter API V2 that was provided through Twitter Developer platform after completion of an application process. This allowed the code to access historic Twitter data. Below are the time points at which tweets were collected.

Table 1: Key Events Timeline for the Lil Nas X Montero Era

Date	Event
Jan 5th, 2021	Release of children's book "C is for Country"
Feb 27th, 2021	Super Bowl LV Commercial
Mar 26th, 2021	Release of Single "Montero (Call me by Your Name)"
Mar 29th, 2021	Promotion of "Satan Shoes"
May 21st, 2021	Release of single "Sun goes Down"
May 22nd, 2021	Saturday Night Live wardrobe malfunction
Jun 27th, 2021	On stage kiss at the BET awards
Jun 28th, 2021	Release of single "Industry Baby"
Sep 15th, 2021	Named among Time's Top 100
Sep 17th, 2021	Release of the album "Montero" and single "That's What I want"

It is important to note that although tweets were collected on these specific dates, they were also collected during time periods around those dates (+/- 5 days), in addition to some random points in time to allow for a comparison between tweets as a reaction to a specific events and general non-event-specific tweets about Lil Nas.

This program was inspired by multiple programs that were already developed for previous research (Edward, 2021) and added new aspects which served the collection of tweets of interest.

The final structure of the data collection program was as follows:

- Twitter API Authentication
- Header definitions
- Definition of all functions to be used during the tweet collection process (such as ***append_to_users*** which adds the data retrieved to a user-specific csv file, and ***connect_to_endpoint*** which requests the tweets from twitter and retrieves them in a json file)
- Keyword definition: this is an important part of the query which allows the retrieval of only tweets that include the items in this keyword. The keyword used in our code was:

  keyword = '(lil nas x -is:retweet OR LIL NAS X -is:retweet OR Lil Nas X -is:retweet OR lil nas -is:retweet OR MONTERO lil nas -is:retweet OR montero lil nas -is:retweet OR CMBYN -is:retweet OR LilNas -is:retweet OR lil nas -is:retweet) lang:en'

  It is worthy to note that the initial direction of work meant to add a *has:geo* constraint as well which would have limited the tweets to those with geolocation enabled, however that led to a very insufficient number of data points and therefore, location was collected using the user's profile information rather than the tweet.
- Definition of the query parameters that decide what values we want returned from each tweet object
- Using a nested while loop, requests were sent to the Twitter API to collect tweets for the query set and between the time periods indicated.
- Start_date and end_date of the search were automatically being updated using the addition of a delta component to the code
- Tokens were being run through using an *if next token* exists loop which provided a bigger volume of tweets by reading tweets from more than just one page of results for each specific constraint statement (query)

The outcome of the data collection program was as follows:
The above detailed data collection procedure led to the collection of 3 different files.

The first file (*tweets.csv*) included over 1.5 million tweets. After removal of duplicates and non-English tweets the data set was left with around 700,000 unique tweets in this comprised of the following:

[tweet_id, author_id, created_at, geo, tweet_text, lang]

The second file (*users.csv*) includes around 420,000 unique user information delineated as:

[user_id, username, user_bio, account_date, user_location]

The third file (*places.csv*) includes exact geotags for 11,600 of the unique tweets and its data storage format is as follows:

[place_name, geo_id, place_type, geo_location]

Because not all the tweets had geolocations attributed to them, the user location was collected as part of the users.csv file in order for it to be later decoded and for a location to be specified for each tweet. The number of tweets collected that had an actual geotag was 11,600. This is around 2% of the entire set of tweets collected. Having the user_location from user profile decoded using GeoPy library (to be explained in the pre-processing section) raises this number of locations from a mere 2% to around 32%. Although most users have a location attributed, Twitter location allows for manual editing, and many are therefore not accurate representations of locations hence making reaching a higher level of location accuracy rather infeasible.

One final thing to note is that to avoid redundancy in data and make the collection process more effective, the -is:retweet parameter is added to the search query thus preventing the collection of retweets of the same tweet repetitively.

B. Data Pre-Processing

Having three different data files made the data collection process more efficient and the code less complicated during running. However, to process the data properly and drive valuable insights, the data in all three files needed cleaning and organising before finally merging it into one data frame on which the analysis and models are performed.

Cleaning the Tweets File

Although the data collection code made it very difficult for any redundancies to exist within the data set of tweets collected, the team was thorough enough to approach the data set and remove duplicates as a first step. These duplicates were removed based on the tweet_id column, therefore keeping only unique tweets. The unique set of tweets was confirmed using the .nunique() call.

Upon ensuring that only unique tweets remain, the next step was to confirm tweet language. Although the request query restricted tweets to English only (using the lang:en) parameter, Twitter API sometimes collects tweets in other languages as well. Therefore, the data set was updated by removing all tweets with an identified language as anything other than English.

Now that the tweets in the data set are those that are relevant to our study, these tweets are further pre-processed to parse all unusual entities retrieved in HTML format to standard format. This is done with reference to Mondal, 2020 which gives detailed insight into the use of the HTML parser using Python. Using the *cleaner()* function created, the text in each tweet was cleaned by:

- Removing @ signs and the user handles from the tweet text
- Removing links (http and others) from tweet text
- Removing hashtag signs but keeping the text that follows as it will provide insight into what users are interested in and what they're tweeting about

Tweet text, upon completion of cleaning, are then are transformed into lower case and all special characters are removed (such as * ? , \). Finally, date of tweet was changed from ISO 8601 standard to a python date time object for all tweets. This concludes the cleaning process for the tweets file.

Cleaning the Users File

The user data collected will be eventually merged with the tweets data and the places data in a comprehensive data frame to be analysed. However, individually cleaning the data set is critical to avoid any redundancy or missing data during the merge. Therefore, an extensive cleaning procedure was performed on the user data file as well.

The data collection code was programmed to collect tweets from distinct users to avoid data leakage and having the same individual's opinion retracted more than once. However, confirming the unique number of users is a must in the data cleaning process and therefore duplicates were removed as a first step.

Upon eliminating duplicates, the GeoPy library was utilized to determine the location (or most approximate location) of all the users in the data set. Given the location name collected with each tweet, the GeoPy library was used to determine the latitude and longitude of each of these locations. After that, a function was created to determine the state and country for each of these lats and longs. All this data was stored in the data set and will be combined with the rest of the data. It is important to note that given Twitter's privacy and user data protection policies, not all user locations were able to be retrieved and some remain under the label none.

Cleaning the Places File

The places file contains the data set that has all geographic location data that was retrieved from accounts that had their geolocation tags open to public. This information serves as reference for other user locations that were retrieved using GeoPy. The only pre-processing needed on this file was the removal of duplicates.

To summarize the data cleaning process undertaken, below is a flowchart with a holistic overview:

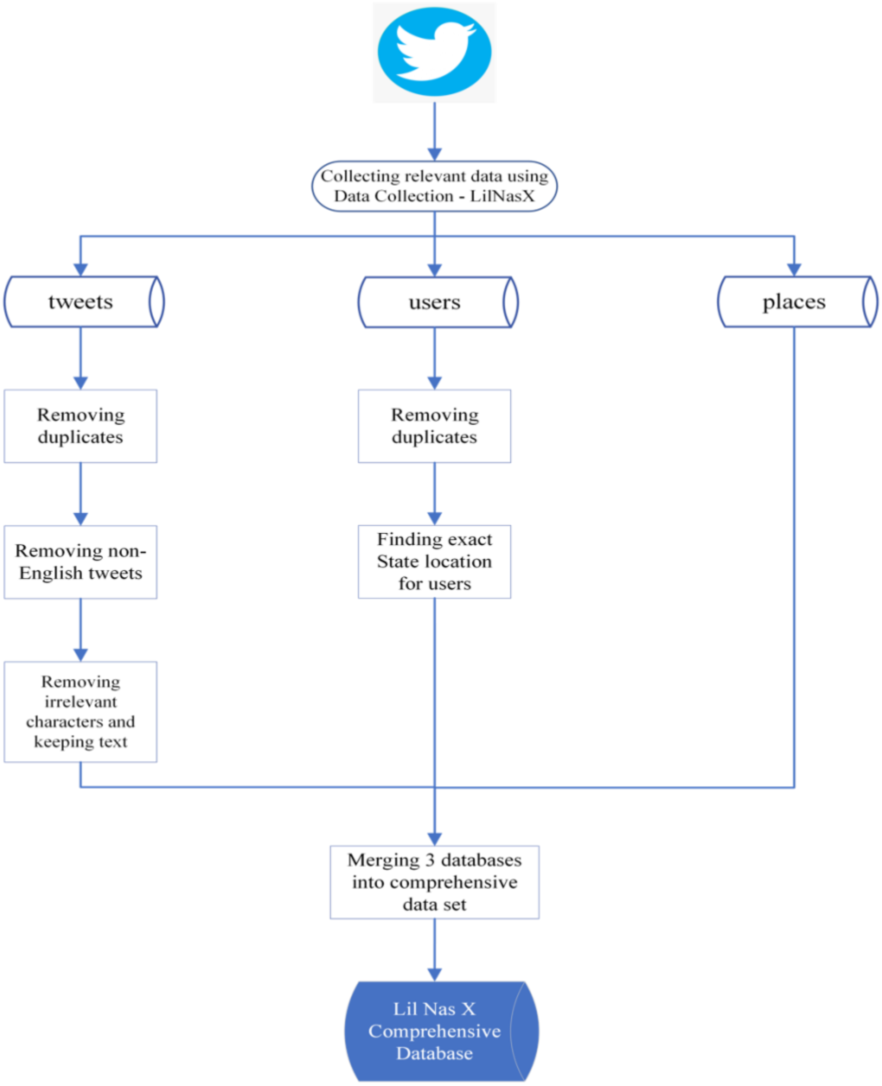

Figure 2: Process Flow Chart for the Data Cleaning Process

C. *Exploratory Data Analysis*

This section provides a rather introductory exploratory analysis into the data prior to classification using the models of choice in later sections. The aim is to get an understanding of the data collected and drive experiment opportunities for the study.

The first step of the exploratory data analysis entailed studying the demography of the data set, most importantly the users and where they are located around the world. Given that we had collected the data about location from user profile, the GeoPy library was used in order to retrieve the latitudes and longitudes, and the GeoPandas library was used to plot a basic map to provide insight into where the majority of the tweets come from.

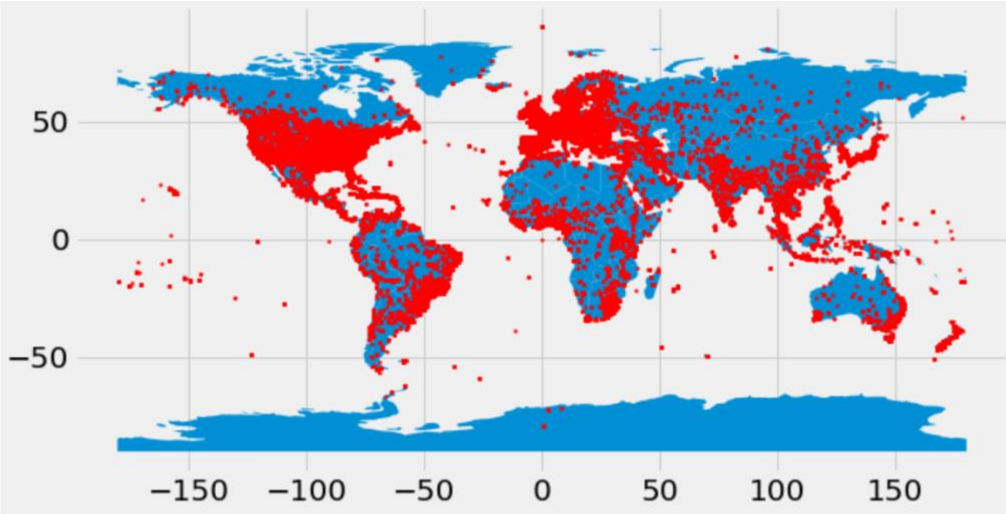

Figure 3: Map Showing Distribution of Tweets Collected Geographically

The map above presents the regions at which there was a lot of twitter activity mentioning Lil Nas X and his work during between March and November 2021. As observed, most of the conversation about Lil Nas X takes place in North America, Central America, Northern Europe, Southern Europe, Western Europe, the Middle East, and Southeastern Asia. There is barely any conversation about the topic in Eastern Europe, Central Asia, Eastern Asia, and Australia. Therefore, the focus of the analysis will lead to conclusion mostly pertaining to the regions of the world that actually were involved in the conversation surrounding Lil Nas X. It is important to note that certain tweets were located to be at random parts of the oceans and uninhabitable areas. These tweets are mostly likely to be the results of users resorting to VPNs on their machines. For the sake of accuracy in this study, these users' locations will be assigned a – none – value.

In order to further focus this research, the countries with the highest numbers of tweets relating to Lil Nas X are also presented.

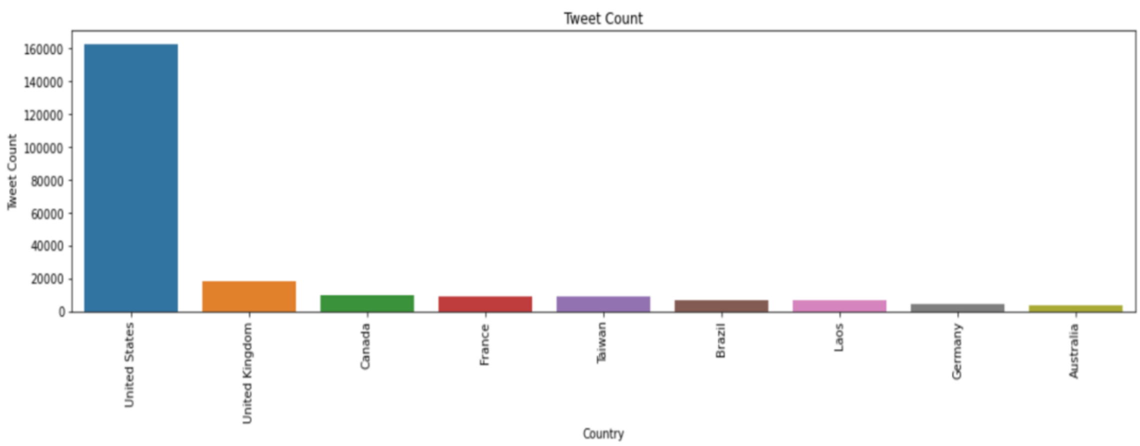

Figure 4: Top 9 Countries in the Discussion on Lil Nas X

IV. Sentiment Analysis Using Vader Sentiment Intensity Analyzer

In aim of reaching a deeper understanding of the conversation surrounding Lil Nas X's Montero era, the attitude and emotions of the tweets published during this era needed to be analyzed. To fulfil that aim, sentiment analysis was performed on all 650,000 tweets collected. A discussion on the results of this analysis will be performed on the general level across all tweets, in addition to more specific levels such as country-level and state-level analysis.

Vader Sentiment Intensity Analyzer was the model of choice in this case for several reasons. Valence Aware Dictionary for Sentiment Reasoning 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 for a positive label on this statement. In addition to that, and although the pre-processing of the tweets prior to performing the sentiment analysis was thorough, VADER also accounts for the meaning behind capitalization, punctuation, and repetition, which makes it the more appropriate model to use when analyzing tweets. This means that if a user tweeted "I HATEEEE Lil Nas X's new video,' their sentiment would be understood to be more negative than a user who tweeted "I hate Lil Nas X's new video."

In our study, VADER Sentiment Intensity Analyzer developed by NLTK (Natural Language Toolkit) was implemented as all literature suggested that it has significantly high accuracy of classification/analysis. The full implementation is found in the *SentimentAnalysis-LilNasX* Jupyter notebook.

There were several steps that were completed to guarantee successful application of the model. Stop words, which are ignorable auxiliary and non-informative words, are also removed by downloading a list of them from nltk.corpus. In addition to that, stemming was performed, which entailed the transformation of all similar words into their essential form (i.e., player and played would be substituted by play). Finally, the NLTK word tokenizer was used to tokenize the tweets into lists of words rather than full statements for the analysis to be accurate.

The below figure shows the distribution of the output data from the Vader Sentiment Intensity Analyzer. The data has been rounded to the nearest 1 and therefore a majority of the data ranging between -0.5 and 0.5 are seen to be at the 0-intensity level. As observed in the bar plot, there are around 60% of the tweets with a neutral intensity attributed to them. However, we can also observe that there is a are considerable tweets with positive and negative sentiment as well. The compound sentiment score presents the compounded values of negative, neutral, and positive sentiments. Therefore, in order to get more insight, the positive sentiment scores and negative sentiment scores are analyzed individually.

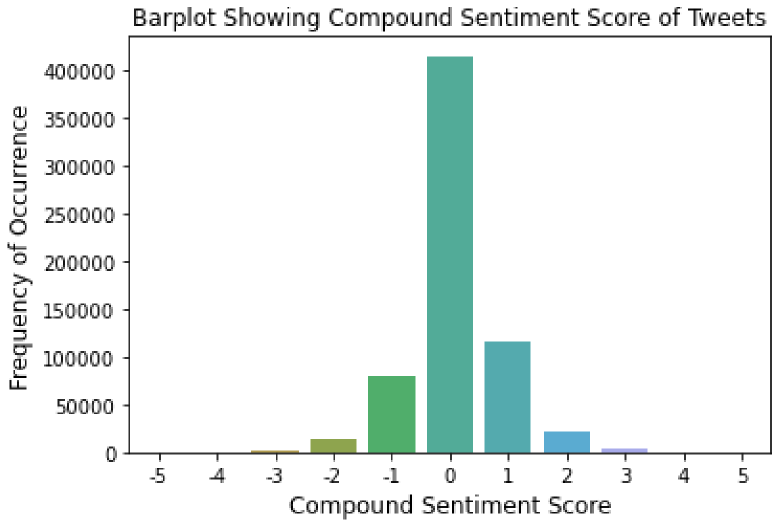

Figure 5: Bar Plot Showing Compound Sentiment Scores of All Tweets

The below graph represents the negative sentiment scores for each of the tweets analyzed. It is observed that 50% of these tweets has some sort of negative sentiment. This is a value-adding insight realized by analyzing one dimension from the overall sentiment intensity analysis. Almost half of the discussion on twitter regarding Lil Nas X's Montero era being of negative sentiment is an indicator of the negative reaction that it triggered. Note that this sentiment analysis does not allow us to conclude that this negativity was targeted at Lil Nas because a good proportion of the tweets showcases Lil Nas supporters reacting to his non-supporters and his critics. The 370,000 tweets classified as neutral have a possibility of being neither hateful nor supportive of Lil Nas X's behavior. However, the possibility for having negative sentiment without using certain keywords associated with negativity and/or hate does exist. Further analysis will be conducted to fill this gap.

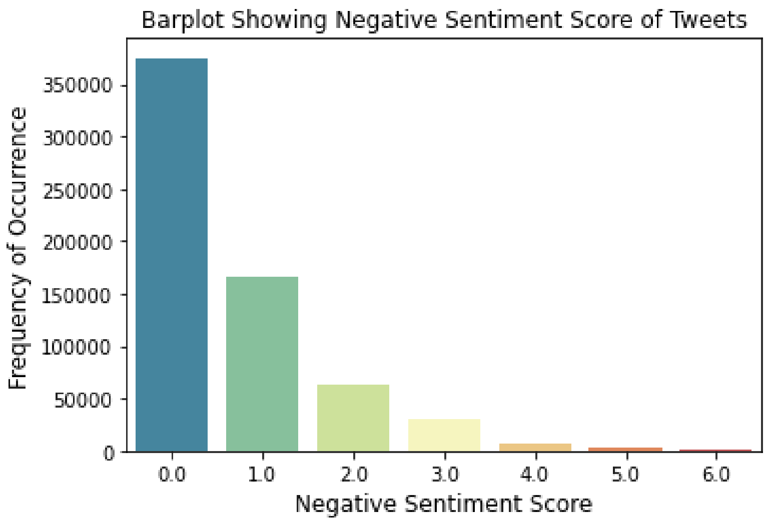

Figure 6: Bar Plot Showing Only Negative Sentiment Scores of All Tweets

One other way to look at what the sentiment scores provide in terms of insight is presented in the adjacent table in which the negative sentiment scores of 0 were removed. We can see that amongst tweets that did contain negative sentiment, the negative intensity has an average score of 29.65. This high value further confirms that the tweets with negative emotion in them included severe/extreme negativity. However, this negativity cannot be labeled as hateful or not prior to determining whether the tweet could be labeled as such. This is covered in the hate speech classifier in a later section.

	unique_values
count	35.000000
mean	29.657143
std	46.197111
min	1.000000
25%	9.500000
50%	18.000000
75%	29.000000
max	270.000000

Table 2: Negative Sentiment Scores Excluding 0-scores

V. <u>Hate Tweets Classification Using XLNet:</u>

A. Model Training & Testing

Based on previous research on speech analysis and labelling (Zhang 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 2020) Beginning with the code provided by Bill Huang of Medium, (2021) we designed a model to label the text of a tweet as either hate speech or not hate speech. This was done with XLNet-Base System which was chosen based on hardware constraints but also found to return acceptable results. Finetuning was completed with the labelled dataset of 24,000 tweets written in 2017 labelled as hate, offensive, or not hate. (Davidson 2017)

As our goal was to differentiate hate vs not hate, we converted all offensive encodings to not hate. Initial runs of training led to validation accuracy ratings between 95% - 99%. However, this was quickly seen to be a result of a large data imbalance. Comparatively, precision and recall within the data labelled “Hate” were half that of the model as a whole. In the original dataset, 6.7% of the tweets were labelled as “Hate” leaving 93.3.% as “Not Hate”. After trial and error of different levels of over and under sampling, we found optimum results occurred with under sampling to create 14.29% level of hate leaving 85.71% non-hate data. Most would still label this as an imbalanced dataset. However, we found increased sample manipulation led much higher rates of overtraining and misclassification.

We then began to manually review the results and found many examples where clear hate speech was mislabeled. Within the test data, we found many instances where clearly identifiable hate speech, particularly regarding the LGBTQ+ and mentally handicapped communities, were not marked as hate. This highlights problems with relying on crowd sourced data regarding contentious topics. As seen in previous work, (Mehrabi 2021) problematic human activity that machine learning models are often trained on easily become replicated in that model’s output.

Within collection scopes of uncontentious topics such as a simple image content classifier, a crowd-based error is less consequential. However, when trying to identify hate speech within a population, the biases that lead a population to participate and accept hate speech will inevitably present themselves in the labelled data.

Based on this realization, we then manually inspected the model's labels against the dataset and reclassified incorrectly labelled text as they were revealed. Several iterative runs of this approach ultimately lead to the results seen in the table below after a 70-30 test-validation split with an overall accuracy rating of 99%.

Table 3: Prelabeled Data Validation Set

Class Specifics				
	Precision	Recall	F1-Score	Support
Hate	0.96	0.90	0.93	542
Not Hate	0.99	0.99	0.99	6938

B. Classification of Data

After running this model over the 650,000 tweets, we then then randomly selected 600 tweets and manually classified them resulting in the following table. This resulted in an overall accuracy of 93.79% +/- 4% at a 95% confidence level.

Table 4: Lil Nas Tweets Classification

Class Specifics				
	Precision	Recall	F1-Score	Support
Hate	0.49	0.80	0.61	37
Not Hate	0.99	0.95	0.94	562

The drop in metrics is not entirely surprising given the fact that our model was trained off an entirely different source of tweets as compared to our own. Both were of course focused on the same goal but understandably 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. Furthermore, 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 is discussing 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. In other cases, the error was from a user actively participating in the reappropriation of a hateful term. This process in which a word that historically would be classified as a slur is instead reappropriated 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. That acceptance is also often dependent on who is saying it, information which is not available to us or the model.

These results - while not being ideal – are acceptable and provide insight. The model emphasizes the importance the recall metric in this situation. Misclassifying hate speech as innocuous is dangerous to marginalized communities and we wanted to ensure that as much as possible could be identified.

VI. Frequent Pattern Growth Implementation on Classified Tweets

A. FP-Growth Implementation and Analysis

Upon data classification as either hate or not hate, we began to perform a more in-depth analysis of the hate being direct at Lil Nas. Appendix A shows results from FPGrowth frequent pattern mining algorithm over tweets labeled as hate in the dataset. Before this was generated, stop words and punctuation were removed along with any keyword used during the original data collection process. It can be assumed that virtually every item in these lists is also frequent with variations of "Lil Nas X".

The results show 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" refers to the rapper "Booser 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 very 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 trainers which included a pentagram charm, a depiction of a devil 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. This allows us to reach the conclusion that the public was in fact in control of the conversation surrounding Lil Nas X. Although the artist himself did engage in Twitter debates and propaganda, it was the users discussing his art and products that were setting the trends and spreading the hateful content – whether in support of him or against him. Several notable association rules built from these frequent patterns are provided below.

Table 5: Association Rules Generated from Frequent Patterns

Terms	Confidence
Shoes => Nike	65.6%
"Satan Shoes" => Nike	65.1%
Satan => Nike	66.3%
Blood => Nike	53.7%
Satan, Shoes => Nike	72.4%

B. FP-Growth Implementation and Analysis

On the data set of classified tweets, and in order to complement and confirm the output of the FP-Growth frequent patters, the tweets classified as hateful were preprocessed and a word cloud was generated as observed in the figure below.

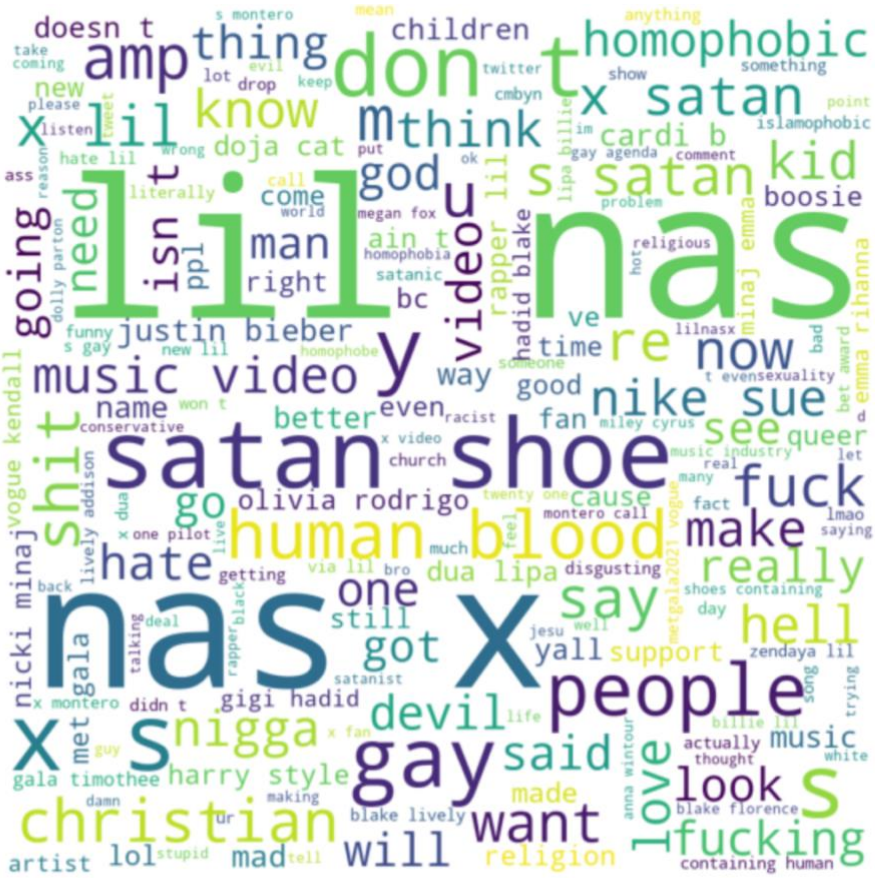

Figure 7: Word Cloud Based on All Hate Tweets

The above word cloud 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 much prevalent.

VII. Analysis & Discussion:

In this section, analyses of different types are to be presented. This analysis is conducted on the final data set of tweets and users which includes the output of the VADER Sentiment Intensity Analyzer and the XLNet classification model.

A. Relationship Between Sentiment Scores and Hate Tweets

Looking into the results of classification and sentiment analysis, we observe the below figure:

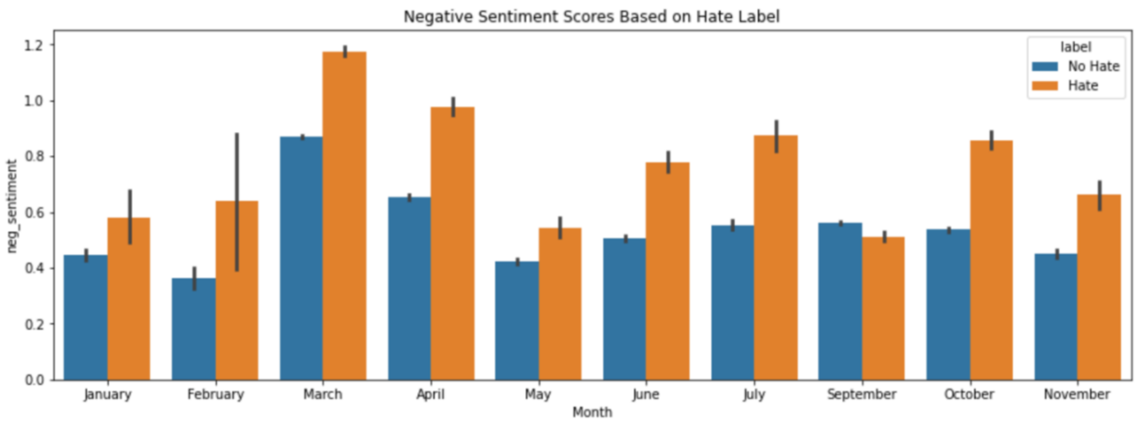

Figure 8: Bar plot Showing Negative Sentiment Based on Hate Label

We can see that the majority of the tweets analyzed are hateful tweets, with the exception of September where we have an almost-equal magnitude of hate vs no hate. It is also observed that hateful tweets have a higher negative sentiment than non-hateful tweets – showing that hateful tweets have more negative sentiment attributed to them. We also observe consistency in the magnitude of negative sentiment across all months. The variability observed is in the hateful tweets whose negative sentiment increases significantly in the March, April, June, July, and October. This is an indicator that Lil Nas X has in fact triggered hateful conversation in these months more than others.

This hateful conversation shows that the public reacted negatively to his art. This 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.

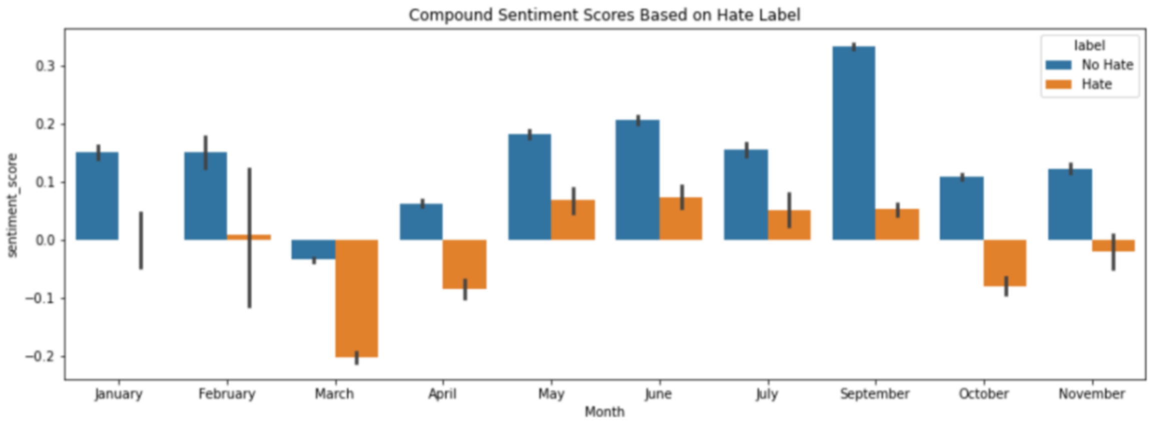

Figure 9: Compound Sentiment Scores Based on Hate Label

One interesting element of conclusion in the above plot is that we observe hate tweets with positive compound sentiment over the course of more than one month. This is an indicator that even though the tweets do indicate hate, they are not necessarily of negative sentiment which means that users behind these tweets were possibly defending Lil Nas X's music videos and/or statements. Another conclusion is that these tweets were all part of the general conversation about the importance of what Lil Nas X is putting out there and discussions without hate attributed to them.

B. Analysis of Hate Tweets around Race, Sexual Orientation, Disability and Religion

The figure below shows the overall proportion of hate related tweets on any given day. This value had an overall range of 15% and mostly stayed just under 10% per day. Some of its peaks were found in late March of 2021 to early April 2021 surrounding the release of the single "Montero" and the controversy of the "Satan Shoes". However, the highest proportion of hate occurred on July 3rd at 17.1% the day after Lil Nas performed at the BET award show where his performance very publicly displayed expressed his sexual orientation.

Proportion of Hate Tweets per Day	
mean	9.392348
std	2.371199
min	4.029423
25%	7.681183
50%	9.155466
75%	10.870629
max	17.056396

Figure 10: Statistics on the Proportion of Hate Tweets per Day

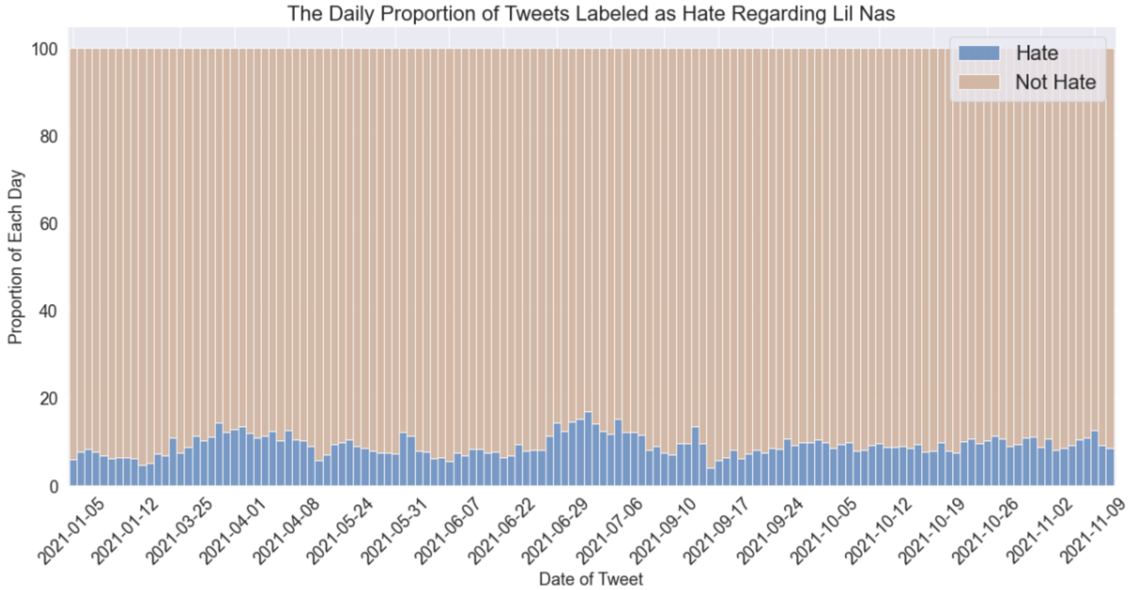

Figure 11: The Proportion of Daily Tweets Marked as Hate

Looking more closely at what constituted each day's hate, we then began to categorize the hate to better understand it's composition. Given the focus of this study is on Lil Nas, a black gay pop rap artist, we chose race, sexual orientation, disability, and religion. All other hate would then be classified as 'Other'. To do this, we compiled a list of hate related slurs surrounding these categories and searched through the tweets to indicate which category they fall into.

Table 6: Proportion of Tweets Using Explicit Slurs Against Minority Groups

Category	Individual Count	Percent of Total Hate Tweets
Race	3,134	4.90%
Sexual Orientation	437	0.68%
Disability	338	0.53%
Religion	541	0.85%
Total	63,987	6.96%

As we can see, despite taking every English slur available from multiple sources, this method only categorized 6.96% of the hate. This result speaks to the work that has been done by Twitter to contain the volume of hate on its platform. Users that repeatedly use these terms are given bans for as short as 12 hours to complete exclusion from the site. The effect of this strategy has been very tangible on the language used withing the platform.

However, hateful slurs are far from the only type of hate speech. The challenge is detecting hate speech when someone intends to marginalize a group without being explicit in their word choice.

We then broadened our strategy to fully rely on the classification of the model. Since every tweet being categorized is already labeled as "hate", we can then use a wider set of words to assign a topic. With the inclusion of otherwise innocuous words that imply a topic without themselves being offensive, we're then able to drastically improve the proportion of categorized tweets. Examples of such words include "skin" for race related topics or "church" for religious. At this point we then add a new category for violence, indicating if the tweet contained words such as kill, shoot, dead, etc.

Table 7: Proportion of Tweets Using Implicit and Explicit Slurs Against Minority Groups

Category	Individual Count	Percent of Total Hate Tweets
Race	8,378	13.49%
Sexual Orientation	14,646	24.49%
Disability	338	00.53%
Religion	17,354	30.98%
Violence	11,160	19.31%
Total Hate Tweets	63,987	58.67%

Please note that the values in the "Percent of Total Hate Tweets" column do not add up to the total at the bottom. This is because many tweets fell into multiple categories. The full list of combinations can be seen in Appendix B. To account for this, the tweets were recategorized in this order: disability > race > religion > sexual orientation > violence. This was chosen to maintain small group integrity by going from smaller to larger counts with two exceptions. First religiously categorized tweets tended to be from religious groups towards a LGBTQ+ community. We wished to preserve this information to differentiate it from other types of LGBTQ+ hate. Second, violence was placed at the end because in most situations, the call to violence was directed at another group. We wanted to preserve this information by only categorizing a tweet as violent if it failed to be captured in any other group before it.

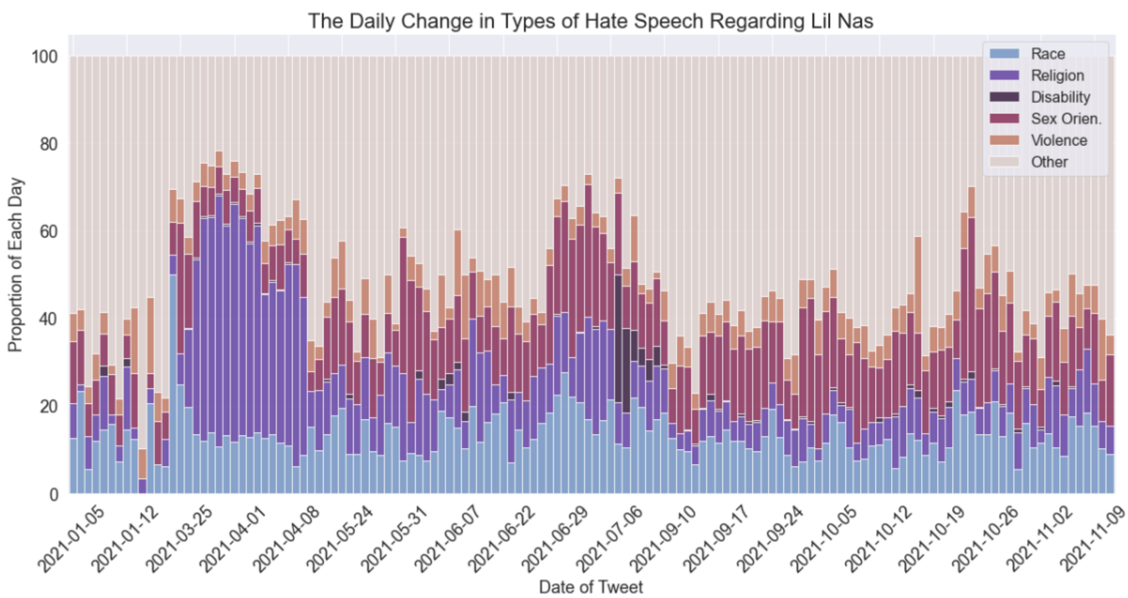

Figure 12: daily change in the types of hate speech through the length of the study

Given the information surrounding what tweets are classified as hate speech, and how those categories are broken down, the figure above plots how each category of hate speech changed on a day-to-day basis. This view allows us to see how the types of hate changed as different events unfolded over the timeline. Distinctly in late March to early April we see 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. This can be seen by the rise of hateful tweets labeled as sexual orientation events, starting in late June, and trailing off through the end of July. Interestingly in early July, there is a spike in disability related hate speech. Most of this is an influx of people using a variation of slurs related to intellectual disabilities. It's not known exactly why this occurred at this moment as there are no major events or news stories connected to Lil Nas that we would have expected to garner that type of attention. This development would be a good point of further research to better understand why the surge in this type of hate speech occurred.

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 completely unconnected to each other, the timing of the events show how positive and negative reactions to his work often end up drawing the attention of people who wish to say hateful things either way.

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. According to the below figure, is a spike in hate related speech corresponding to the release of the album. However, no category of hate dominates the conversation as others had previously.

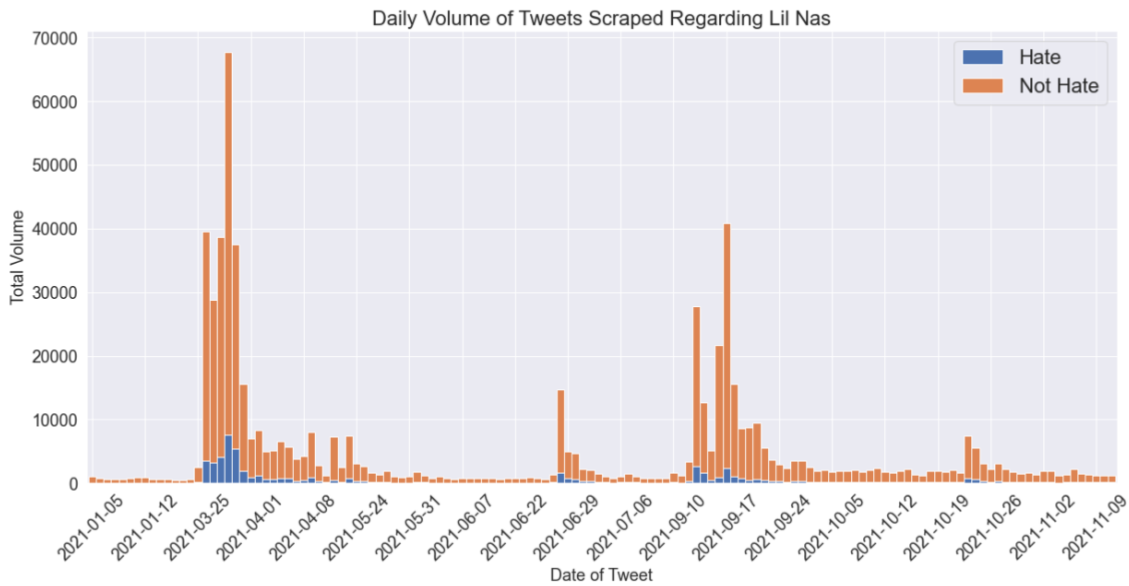

Figure 13:volume of tweets per day mentioning Lil Nas

C. Location-Based Analysis

Based on the observations made in the exploratory data analysis section, location-based analysis was performed in order to study reactions across different nations and parts of the world.

The figure below shows the top 5 countries with the highest percentages of hateful tweets during the Lil Nas X Montero Era.

Country	Percent
United States	59.035971
United Kingdom	5.859220
Canada	3.384205
France	3.016199
Taiwan	2.659018

Figure 14: Countries with the Highest Hate Speech Percentage

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. On the right side are word clouds with a comprehensive analysis of most frequent words used by users of these countries. On the right side are word clouds with words used by tweets classified as hateful only.

Figure 15: Word Cloud from USA Non-Hate Tweets

Figure 16: Word Cloud from USA Hate Tweets

In the case of the United States of America, it is evident that the hateful discussion surrounding Lil Nas X had three major components: race, sexual orientation, and religion. We see words such as “Christian”, “Satan”, “God”, “devil”, and “hell” being used most frequently and that shows that the US community reacted hatefully because Lil Nas has used religious insinuations in his first music video. A lot of mentioning of kids takes place which drives us to conclude that American do not want their kids being exposed to such content. Homophobia is present in different forms of the word, which is an indicator of people spreading hate against homophobes. Finally, the ‘N’ word is utilized in many spaces, in addition to the word black which shows that the US demographic is yet to stop labelling people based on their race.

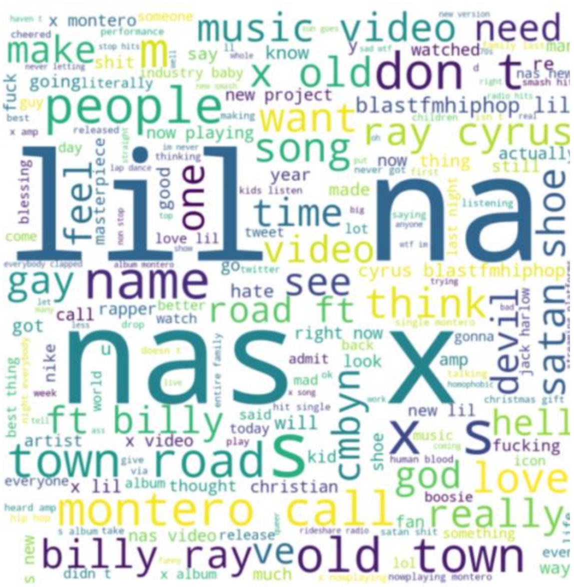

Figure 17: Word Cloud from UK Non-Hate Tweets

Figure 18: Word Cloud from UK Hate Tweet

Similar to the case of the United States, the conversation across the United Kingdom was also focused on sexuality and religion. However, we see that the words used when discussing sexuality are more progressive words than those used in the U.S. such as the dominance of the word “queer” rather than gay like in the States. The conversation does also include Christianity and the church.

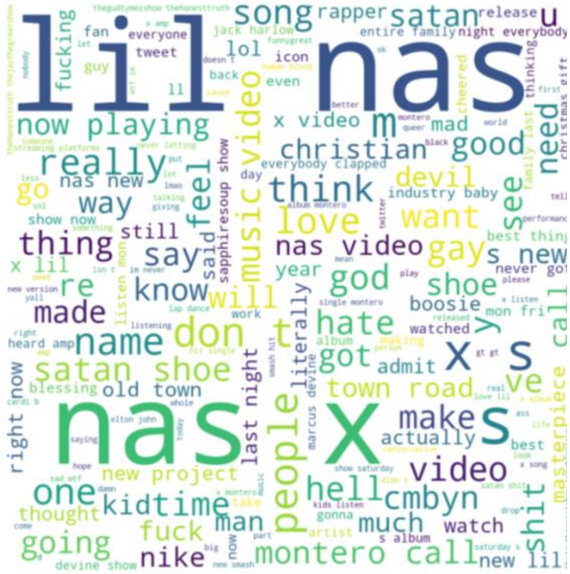

Figure 19: Word Cloud from Canada Non-Hate Tweets

Figure 20: Word Cloud from Canada Hate Tweets

In Canada, the conversation was similar to that in the U.S. but did not include much focus on religion (doesn’t mean religious tweets are none). There is a lot focus and repetition of the terms “human” and “support” which indicate a positive reaction from the public to Lil Nas X’s montero era but hate was projected towards those who did not agree with that.

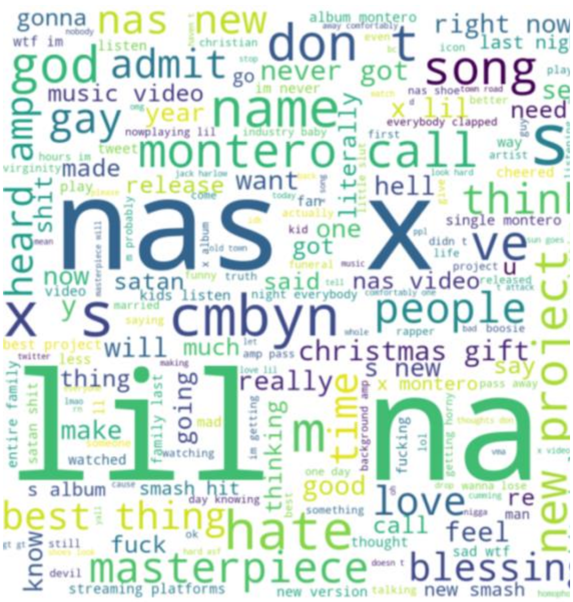

Figure 21: Word Cloud from France Non-Hate Tweets

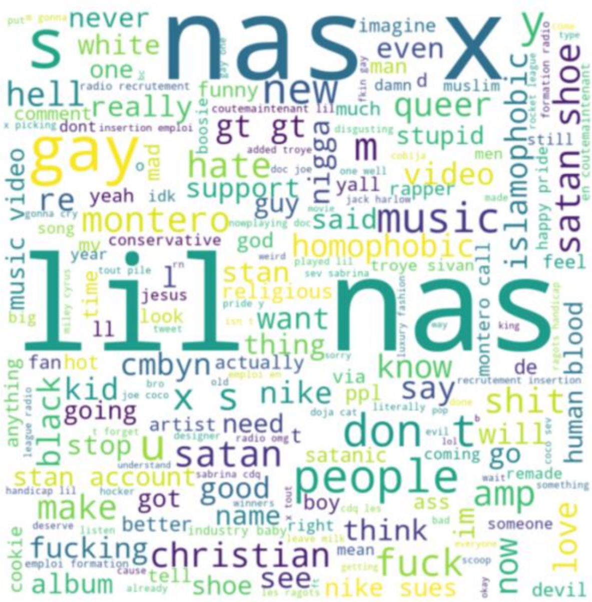

Figure 22: Word Cloud from France Hate Tweets

An interesting term in the French cloud is “islamophobia”. The French were criticizing Lil Nas because of a tweet of his at the age of 16 being Islamophobic. This shows that hate against him in France was because of the impression that he was an Islamophobe, and not because of his sexuality and race.

Figure 23: Word Cloud from Taiwan Non-Hate Tweets

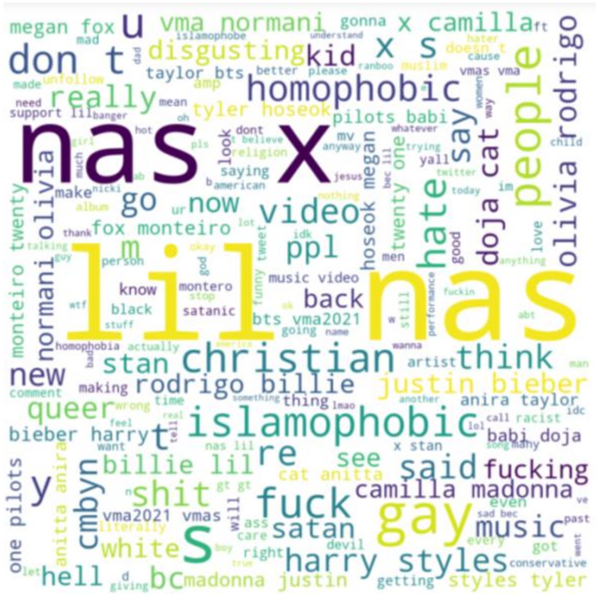

Figure 24: Word Cloud from Taiwan Hate Tweets

The final word cloud is that of Taiwan, a country with a very different structure than all of those above. Islamophobia was also a word of frequent use. We can observe a lot of conversation centered around homophobia, disgust, and queerness. Furthermore, this is a country where “child” was mentioned more often than others, indicating a concern for children being exposed to Lil Nas X material.

The reason behind generating word clouds for each of these five countries, other than the fact that they have the largest share of hate tweets attributed to them, is the fact that they all lie in different geographic regions and have different cultures, trends, and perspectives when it comes to homophobia, racism, and tolerance.

Some insights from the above are:

- 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 country.
- 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 the one being held in North America (US and Canada) and shows the focus of these different communities.

VIII. Conclusion

After analyzing tweets of over 600,000 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 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. The world wishes to believe that the United States, with all the progress it has made on the fronts of human rights and equality, is not still home to such a magnitude of racism and homophobia. However, the analysis shows that unfortunately, controversial art still triggers a large number of individuals – and their criticism is not solely on the sexual orientation of this artist, but they automatically drive the conversation into a racist one. Interestingly enough, countries like 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.

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.

IX. References

[1] *Twitter: Most users by country*. Statista. (2021, November 19). Retrieved November 30, 2021, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[2] Angelo, P. J., & Bocci, D. (2021, January 29). *The changing landscape of Global LGBTQ+ rights*. Council on Foreign Relations. Retrieved December 2, 2021, from https://www.cfr.org/article/changing-landscape-global-lgbtq-rights?gclid=Cj0KCQiA15yNBhDTARIsAGnwe0Wltd74MlSYk2rIxg633xlIWYXA0C2KmKuX4WRasds2aTj9zy82m54aAlE-EALw_wcB#chapter-title-0-3.

[3] Litchfield, C., Kavanagh, E., Osborne, J., & Jones, I. (2018). Social media and the politics of gender, race and identity: The case of Serena Williams. *European Journal for Sport and Society*, *15*(2), 154–170. https://doi.org/10.1080/16138171.2018.1452870

[4] Edward, A. (2021, June 17). *An extensive guide to collecting tweets from Twitter API V2 for academic research using python 3*. Medium. Retrieved from https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a.

[5] Mondal, S. (2020, November 22). *Twitter data cleaning and preprocessing for Data Science*. Medium. Retrieved from https://medium.com/swlh/twitter-data-cleaning-and-preprocessing-for-data-science-3ca0ea80e5cd.

[6] Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[7] Beri, A. (2020, May 27). *Sentimental analysis using vader*. Medium. Retrieved from https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664.

[6] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan A., (July 2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput*. Surv. 54, 6, Article 115 35 pages. https://doi.org/10.1145/3457607

[7] Huang B. (2019, August 5). Text Classification with XLNet in Action. *Medium*. Retrieved from https://medium.com/@yingbiao/text-classification-with-xlnet-in-action-869029246f7e

[8] Coles, G. (2016). The Exorcism of Language: Reclaimed Derogatory Terms and Their Limits. *College English*, 78(5), 424–446. http://www.jstor.org/stable/44075135

[9] Davidson, T., Warmsley, D., Macy, M., Weber. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, 512 – 515.

[10] Schmidt, A and Wiegand, M. (April 2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1-10. Association for Computational Linguistics, https://aclanthology.org/W17-1101

[11] Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J (2021). Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study. *JMIR Infodemiology*. https://doi.org/10.2196/26769

[12] Mutanga, R. T., Naicker, N., & O, O. (2020). Hate speech detection in Twitter using Transformer methods. *International Journal of Advanced Computer Science and Applications*, *11*(9). https://doi.org/10.14569/ijacsa.2020.0110972

[13] Matthew L. Williams, Pete Burnap, Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data, *The British Journal of Criminology*, Volume 56, Issue 2, March 2016, Pages 211–238, https://doi.org/10.1093/bjc/azv059

[14] (2021). Hatebase. Retrieved from https://www.hatebase.org.

X. Appendix A: FPGrowth of tweets labeled as “Hate” from the Lil Nas dataset

[('So',), ('never',), ('666',), ('sues',), ('Nike', 'sues'), ('making',),
('Boosie',), ('Gay',), ('care',), ('2',), ('ever',), ('Lipa',), ('Dua',
'Lipa'), ('need',), ('believe',), ('Dua',), ('also',), ('Not',), ('artist',),
('us',), ('listen',), ('Megan',), ('can’t',), ('1',), ('Rapper',), ('real',),
('fan', 'music'), ('I’m', 'fan', 'music'), ('I’m', 'fan'), ('I’m', 'I’m',
'fan'), ('In',), ('MONTERO',), ('That',), ('ppl',), ('His',), ('hes',),
('gays',), ('still',), ('To',), ('Im',), ('What',), ('Is',), ('better',),
('Bieber',), ('Bieber', 'Justin'), ('THE',), ('makes',), ('Cardi',), ('B',
'Cardi'), ('Black',), ('Human',), ('Blood', 'Human'), ('B',), ('Olivia',),
('nigga',), ('Montero',), ('satanic',), ('way',), ('Justin',),
('Christians',), ('made',), ('Blood',), ('Doja',), ('He',), ('If',),
('thing',), ('would',), ('see',), ('ass',), ('good',), ('And',), ('devil',),
('white',), ('support',), ('Nike', 'Shoes’'), ('Nike', 'Shoes’', '‘Satan'),
('Shoes’', '‘Satan'), ('make',), ('he’s',), ('love',), ('I', 'love'),
('‘Satan',), ('Nike', '‘Satan'), ('mad',), ('going',), ('kids',), ('it’s',),
('right',), ('go',), ('YouTube',), ('YouTube', 'via'), ('2021',), ('man',),
('A',), ('want',), ('yall',), ('shoes',), ('got',), ('even',), ('fucking',),
('hell',), ('You',), ('satan',), ('said',), ('u',), ('queer',), ('say',),
('Satan', 'blood'), ('Satan', 'blood', 'human'), ('Nike', 'blood'), ('Nike',
'blood', 'human'), ('blood', 'human'), ('Satan', 'human'), ('Nike', 'human'),
('black',), ('homophobic',), ('new',), ('fuck',), ('really',), ('I',
'really'), ('I', 'I’m', 'I’m'), ('I', 'I’m', 'I’m', 'music'), ('I’m', 'I’m',
'music'), ('I’m', 'music'), ('I', 'I’m', 'music'), ('I', 'I’m'), ('get',),
('I', 'get'), ('hate',), ('I', 'hate'), ('know',), ('I', 'know'), ('dont',),
('I', 'dont'), ('one',), ('X’s',), ('think',), ('I', 'think'), ('This',),
('I', 'This'), ('Nike', 'Shoes', 'Xs'), ('Nike', 'Satan', 'Shoes', 'Xs'),
('Nike', 'Satan', 'Xs'), ('Shoes', 'Xs'), ('Satan', 'Shoes', 'Xs'), ('Satan',
'Xs'), ('don’t',), ('I', 'don’t'), ('Nike', 'via'), ('Satan', 'via'),
('shit',), ('I', 'shit'), ('y’all',), ('Nike', 'Shoes'), ('Nike', 'Satan',
'Shoes'), ('Satan', 'Shoes'), ('I', 'amp'), ('amp', 'amp'), ('I', 'video'),
('music', 'video'), ('The',), ('music',), ('I', 'music'), ('gay', 'people'),
('I', 'people'), ('Satan',), ('Nike', 'Satan'), ('Nike',), ('gay', 'gay'),
('I', 'gay'), ('I',), ('I', 'I')]

XI. Appendix B: Hateful tweets categorization

Race	Sex Orin	Disibility	Religion	Violence	Count
F	F	F	F	F	26445
F	F	F	F	T	3199
F	F	F	T	F	9688
F	F	F	T	T	4053
F	F	T	F	F	72
F	F	T	F	T	19
F	F	T	T	F	36
F	F	T	T	T	18
F	T	F	F	F	6868
F	T	F	F	T	1224
F	T	F	T	F	2766
F	T	F	T	T	819
F	T	T	F	F	15
F	T	T	F	T	109
F	T	T	T	F	15
F	T	T	T	T	7
T	F	F	F	F	2426
T	F	F	F	T	1079
T	F	F	T	F	682
T	F	F	T	T	571
T	F	T	F	F	11
T	F	T	F	T	4
T	F	T	T	F	3
T	F	T	T	T	5
T	T	F	F	F	1870
T	T	F	F	T	797
T	T	F	T	F	712
T	T	F	T	T	438
T	T	T	F	F	9
T	T	T	F	T	11
T	T	T	T	F	5
T	T	T	T	T	2