Predicting Hate Crime Biases Using Multiclass Classification Machine Learning Models

Jimmy Hoang
Adviser: Xiaoyan Li

Abstract

In comparison to the widely researched topic of hate speech in machine learning, hate crimes are far less studied, despite being quite similar in nature. Unlike hate speech, though, hate crimes are criminal offenses unprotected by the law, so we can benefit as a society by actively working towards better understanding hate crimes to promote early prevention and effective addressal. This project focuses on the FBI's Hate Crime Statistics dataset that provides incident-level information including dates, locations, offenses, offenders, victims, and biases involved in each reported hate crime since 1991. Utilizing data analysis methodology and multiclass classification models like the Naive Bayes, Logistic Regression, AdaBoost, Random Forest, K-Nearest Neighbors, and Artificial Neural Network classifiers, this project examines trends across hate crimes, explores the possibility of predicting biases involved in hate crimes, and attempts to identify features that have a significant influence on those biases. Using the weighted F1 metric to evaluate model performance, the results of this project highlight that the Artificial Neural Network achieves the highest weighted F1 score of 62% for the classification task and that the offender race, location type, and offense type involved in a hate crime have a considerable bearing on successfully predicting the bias motivating it.

1. Introduction

The FBI defines a hate crime to be a "criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity."[1] According to the Bureau of Justice Statistics's (BJS) National

[1] https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr/hate-crime

Crime Victimization Survey (NCVS), U.S. residents experienced approximately 246,000 hate crime victimizations each year from 2005 to 2019, [2] and they often never reported these incidents to law enforcement. In this society where individual differences should be accepted and embraced, hate crimes have been far too numerous and detrimental, leaving entire communities feeling oppressed and isolated.

1.1. Motivation

The motivation behind my project stems from the prominent issue of hate crimes plaguing the United States, especially in recent years. In fact, the FBI, who tracks bias-motivated attacks, has released data indicating that the number of reported hate crimes rose to the highest level yet in 2021, a 31% increase from the previous peak in 2020 and an 11% increase from the next previous peak in 2001.[3] This growing trend is not surprising, though, as it can arguably be attributed to the surge of targeted hate caused and amplified by major sociopolitical events affecting the entire nation like the 2016 and 2020 presidential elections as well as the COVID-19 global pandemic. Additionally, in this technology-driven society where social media has seeped its way into the different parts of our everyday lives, we have brought the discussion surrounding hate crimes on to the internet, allowing for even more hate groups to collude. In recent years, anti-hate movements advocating for African Americans, the LGBTQIA+ community, Asian and Pacific Islanders, and the Jewish community have been at the forefront of the hate crime discussion, but I would like to expand the discussion by also studying other targeted groups in my work to reflect the fact that hate crimes can affect anyone.

1.2. Goal

The goal of my research is three-fold. First, I wish to identify trends across hate crimes, looking for metrics like the most commonly targeted groups, frequent areas where they occur, the offenses that are most present, and who commonly commits them. Second, I want to investigate the feasibility of predicting the bias motivating a hate crime given the specifics of the incident. Third, I seek to

[2] https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/hcv0519_1.pdf
[3] https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime

pinpoint the factors involved in a hate crime that are most impactful upon those predictions. Hate crime laws in the United States vary widely across jurisdictions and biases, [4] so I hope my work can be useful for law makers and law enforcement. Specifically, I want to shed light on potential indicators of hate crime biases and contribute data that will motivate changes in the way we currently prevent and address hate crimes as well as how we allocate support and resources to communities in need.

2. Background and Related Work

Machine learning literature focused on hate crimes, unrelated to the close yet far more studied relative of hate speech, is scarce. In fact, little research has been done on hate crimes, especially in regards to the task of classification, because of how complex they can be. The research that has been done, though, attempts to further understand hate crimes through yearly trend prediction, spatial prediction, and crime rate prediction. Each of the following works yielded significant results that are paving the way for hate crime research in machine learning.

2.1. American Hate Crime Trends Prediction With Event Extraction

Inspired by previous work that utilized event extraction technologies to identify local hate crimes not included in the FBI's hate crime reports, Han et al. incorporated event-related variables into regression models to improve yearly hate crime prediction [2]. To do so, they first trained an event extraction model on the Patch Hate Crime Dataset to detect the presence of hate crimes in news and achieved approximately 82% precision, recall, and F1 scores. Then, they used the previously trained model to identify hate crimes in their scraped New York Times corpus and extracted two time series: the number of quarterly news reports and events with hate crimes detected by the model. Han et al. then built a baseline regression model using explainable factors related to hate crimes such as "aggravated_assault_rate", "population", and "total_law_enforcement_employees". Finally, they added parameters constructed from the two previously extracted time series into their baseline model, estimated all parameters using maximum likelihood estimation, and predicted the

[4] https://www.justice.gov/hatecrimes/laws-and-policies

FBI's quarterly hate crime report trends. Their augmented model achieved RMSE and MAPE scores that were half of those of the baseline model. Han et al.'s experimental results demonstrated that event-related factors can effectively improve hate crime prediction, proving that their event extraction module was effective.

2.2. Spatial Prediction of Sparse Events Using a Discrete Global Grid System; A Case Study of Hate Crimes in the USA

Motivated by the challenging problem of spatially predicting any geographic phenomenon with many influencing factors, Jendryke and McClure utilized an innovative approach that combines a Discrete Global Grid System (DGGS) with machine learning to analyze spatial hate crime trends [3]. To do so, they first scraped and geocoded hate crime data from the Southern Poverty Law Center (SPLC), demographic data from the National Historical Geographic Information System, and police death data from the Officer Down Memorial Page for their DGGS. Then, they trained an Artificial Neural Network (ANN) on 50% of the SPLC dataset to map and spatially predict hate crimes in the United States using the structured input from their DGGS. After using a Spatial Lag Regression Model to evaluate and rank the outputs of their ANN with spatial considerations, Jendryke and McClure achieved 78% accurate predictions and verified their predictions at the state level against the FBI's Hate Crime Statistics dataset with a fit of 80%. They found that the areas with the highest exposure to hate crimes were those where urban areas transition from higher density to lower density areas and concluded that machine learning can predict hate crimes to some degree of accuracy and precision.

2.3. Hate Crime Analysis Based on Artificial Intelligence Methods

Interested in the state of hate crimes before and after the 2016 U.S. presidential election, Wang utilized a variety of different machine learning tasks to explore the extent to which some factors may lead to higher hate crime rates [4]. Specifically, the factors explored were statistics and demographics relating to a population's education level, income level, racial makeup, citizenship status, and political affiliations. Using a dataset with hate crime rates of states from 2010 to 2016 as

well as the aforementioned features, Wang first performed linear regression on the features to find that a population's "median_household_income" and "share_non_white" features had the strongest correlations to crime rate. Afterwards, K-Means Clustering was used to group hate crime data into five classes according to crime rate levels (0-4), and then K-Nearest Neighbors (KNN) was used to predict those levels for hate crime data. Wang found that training the KNN model with just the "median_household_income" and "share_non_white" features could achieve 50% accuracy for crime level classification and concluded that the increase of hate crimes in 2016 was mainly correlated with income inequality, median household income, and race.

3. Approach

All of the above studies explored hate crimes from different perspectives and were successful in their respective ways, but none looked for the possible insights that may be revealed from specifically studying the biases involved in the hate crimes themselves, the whole reason why they are *hate* crimes. For my approach then, I decided to perform analysis and multiclass classification on the FBI's Hate Crime Statistics dataset in order to further understand hate crimes at the bias level and explore the possibility of predicting biases given the details of a hate crime. To my knowledge, this approach to studying hate crimes has never been done before.

3.1. The Dataset

The U.S. Department of Justice administers two statistical programs to measure the magnitude, nature, and impact of hate crimes in the United States: the FBI's Uniform Crime Reporting (UCR) Program and the BJS's National Crime Victimization Survey (NCVS).[5] While they both provide valuable information about hate crimes in the U.S., they use different methods and focus on different aspects. From a high level, the UCR's Hate Crime Statistics dataset provides a measure of the number and types of crimes reported to law enforcement agencies throughout the country while the NCVS also includes those not reported to law enforcement authorities. Because it is more accessible and digestible for the scope of this research and only includes crimes confirmed as hate crimes by

[5] https://bjs.ojp.gov/content/pub/pdf/ntcm.pdf

law enforcement, only the Hate Crime Statistics dataset [1] will be studied. The dataset contains approximately 225,000 confirmed hate crimes recorded from 1991 to 2021 and provides general information about each hate crime including the date, location, victims, offenders, offenses, and biases involved. While Jendryke and McClure also used the same dataset in their work, I worked with the whole dataset during every step of the machine learning process as opposed to solely using the spatial information provided for model validation.

3.2. The Tasks

The three machine learning and data science tasks explored in this research were chosen with an emphasis on examining hate crime biases to reflect their importance in constituting and diversifying the hate crimes they are associated with. Carrying out these tasks revealed valuable insights about different biases.

1. First, I identified trends across and between hate crime biases through exploratory data analysis.
2. Then, I trained and tuned Naive Bayes (NB), Logistic Regression (LR), AdaBoost (AB), Random Forest (RF), K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) classification models to predict which group a hate crime is targeting.
3. Lastly, I extracted the most important factors in determining the presence of biases in hate crimes.

3.3. The Classification Output

Following the FBI Crime Data Explorer Tool's established method of grouping communities affected by hate crimes, [6] I trained the classification models to predict the following six biases: Anti-Race/Ethnicity/Ancestry, Anti-Religion, Anti-Sexual Orientation, Anti-Disability, Anti-Gender, and Anti-Gender Identity. The raw dataset has 35 unique biases motivating hate crimes, so I had to decrease the complexity of the classification task.

[6] https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/hate-crime

4. Implementation

For the implementation of my project, there were four main stages: data collection and analysis, data preparation, model selection and training, and model evaluation and tuning.

4.1. Data Collection and Analysis

After downloading and importing the Hate Crime Statistics dataset from the FBI's Crime Data Explorer Tool, I first used the pandas, NumPy, and matplotlib libraries to examine its dimensionality and features to gauge what information I had available to work with. The raw dataset contained 226,328 unique hate crimes broken down by 28 features including year, reporting agency, region/division/state, surrounding population density, offender race/ethnicity, offense type, location type, bias description, victim type, and the number of juvenile and adult offenders and victims. The majority of features did not contain missing values, but I did find that that a fifth of the feature columns were only a quarter filled. These pertained to the counts for offenders and victims broken down by age group. I furthermore examined feature frequencies because most of the dataset's features were categorical and discovered that although some feature columns like offender race and offender ethnicity were "complete", they still had many, if not a majority of, values labeled "Unknown", "Not Specified", or "Other". Based on these initial observations, there was a considerable amount of dataset cleaning that needed to be done. Examining feature frequencies with visualizations also revealed interesting insights about the dataset that are outlined and discussed in Section 5.1.

4.2. Data Preparation

After exploratory data analysis, I first removed columns that were repetitive, sparse, or irrelevant. These included columns like state abbreviation that were encodings of others, columns like offender ethnicity that were more than half empty, and columns like incident id that provided uninteresting information. Then, I restricted categorical features to only allow one value per category to decrease dimensionality. For example, an offense type category could have multiple offenses delimited by

semicolons. I also manually removed extremely sparse feature categories like the location type of ATM to ensure that my training and testing sets would have the same feature columns after one-hot encoding all the categorical features of the dataset. Additionally, I removed rows with unknown or multiple biases to streamline classification and relabeled every hate crime's bias description according to its associated overarching target class. Afterwards, I used the scikit-learn library to split my dataset into training and test sets using an 80:20 ratio (80:10:10 for tuning the ANN with an additional validation set) and normalize my numerical features. To complete data preparation, I one-hot encoded the categorical features of my training (and validation) and testing sets and addressed the high imbalance of the 6 target classes. As shown in Figure 1, the dataset is dominated by Anti-Race/Ethnicity/Ancestry hate crimes, so I oversampled every other class to it. Before data

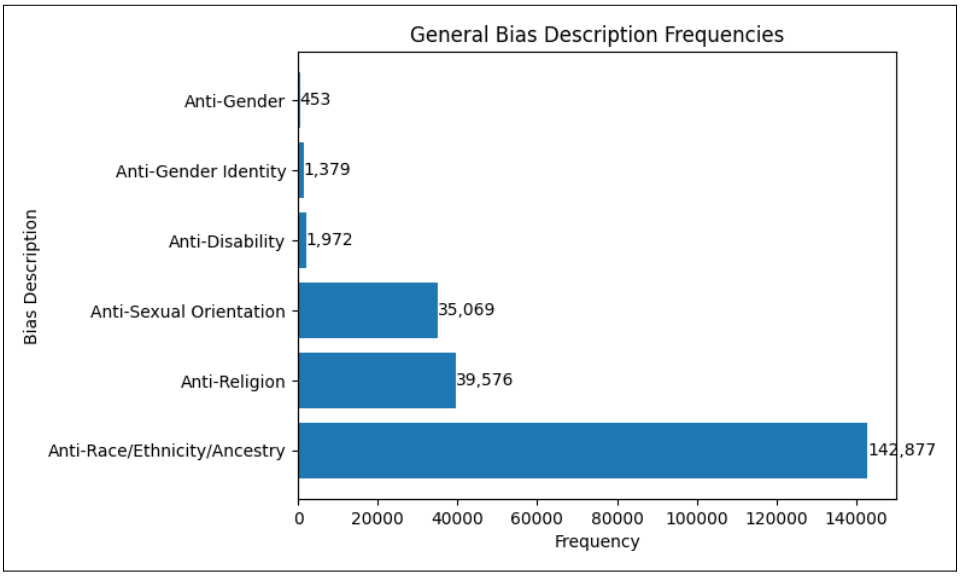


Figure 1: Frequencies of the 6 target classes before resampling.

preparation, the raw dataset had dimensions (226328, 26). Before resampling, my training data had dimensions (177060, 204), and my testing data had dimensions (44266, 204).[7] After resampling, my training data had dimensions (685806, 204), and as a preliminary way to explore feature importance, I used SelectKBest to identify the top 25 features from the training data based on chi-square scores. I discuss them in Section 5.3.

[7]For the ANN, the validation and test sets each had 22133 samples.

4.3. Model Selection and Training

To explore the different capabilities and performances of a classifier on the data, I selected a variety of multiclass classification models to use in my research that are either popular, were used in related works, and/or have unique strengths and weaknesses that can reveal interesting insights about the data. Furthermore, I created and trained all of the models using the scikit-learn library,[8] opting for an 80:20 split.[9] The selected models are outlined below.

4.3.1. Multiclass Classification Models

1. **Dummy** - Found in the scikit-learn library, DummyClassifier's purpose is to serve as a baseline model to compare with more complex models. It ignores the dataset's input features, and its behavior follows a specified strategy which I chose to be "uniform". This means that the classifier generates predictions uniformly at random, so each of the six bias groups have an equal probability of being selected as the final prediction.
2. **Naive Bayes (NB)** - Despite its over-simplified assumptions, this classifier has performed well in real-world situations and is quite fast. It is based on Bayes' Theorem, using the "naive" assumption that there is conditional independence between each pair of predictors so that each predictor independently impacts the output classification. Since the set of features before one-hot encoding was mostly categorical, about $\frac{3}{4}$ of them, I used CategoricalNB, which implements the categorical naive Bayes algorithm for categorically distributed data. It assumes that each feature has its own categorical distribution. Since NB is a very straightforward classifier that is easy to interpret, I used it.
3. **Logistic Regression (LR)** The multinomial version of this classifier measures the relationship between a categorical dependent variable and one or more independent variables. It estimates the optimal weights for each independent variable through optimization techniques like gradient descent to find the best fit of log odds. Once the optimal coefficients are found, the conditional

[8] The keras library was used for the ANN.
[9] 80:10:10 for the ANN model.

probabilities of predicting each class can be calculated. I used this classifier because it is also linear like NB but has lower bias, something worth considering the importance of in evaluation.

4. **AdaBoost (AB)** - As an ensemble method, this classifier uses adaptive boosting to combine the efforts of multiple weak classifiers, usually decision trees, into one strong classifier. Decision trees are structures such that each internal node represents a test for a predictor, each split in the node represents the outcome of the test, and each leaf node represents a class label. A path from the root to a leaf represents classification rules so that the classifier can gather information about the relative difficulty of predicting each sample. This way, future trees can focus on the more difficult samples. I decided to use this classifier because it would provide me with a stronger performance than if I had just used and tuned a single decision tree.
5. **Random Forest (RF)** - This is also an ensemble method, but unlike adaptive boosting, this classifier creates a "forest" of decision trees, and then the model combines all of the decision tree predictions to form its own overarching prediction. I used this classifier because it is very popular in machine learning for its robustness, it is compatible and easy to use with different input feature types, and it also easily computes the importance of input features. Additionally, I wanted to compare RF to the similar ensemble method of AB.
6. **K-Nearest Neighbors (KNN)** - This classifier uses proximity to find the k-nearest points to a data point p and subsequently uses their associated classes to predict p's class. I used this classifier because it assumes that similar points can be found near one another which enables it to handle very complex data like hate crimes.
7. **Artificial Neural Network (ANN)** - This powerful classifier uses interconnected layers of nodes to create an adaptive system that a computer can use to effectively learn on its own. Information is introduced into the ANN through the input layer, processed in hidden layers that compute feature weights, and then outputted as a class prediction. By using learning algorithms that can independently make adjustments and learn as they receive new input, ANNs are an effective tool for non-linear statistical data modeling. I used this classifier because of its powerful ability to recognize patterns in complex data.

4.3.2. Initial Training

In order to record baseline model performances for later comparison with tuned model performances, I initially trained all of the models with their default parameters given by scikit-learn.

4.4. Model Evaluation and Tuning

I evaluated the different models using the following two performance metrics:

1. F1 Score - This metric measures the harmonic mean of precision and recall and is useful when there is an uneven class distribution. The calculation for this metric is given by $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. Classification tasks usually entail accuracy as an evaluation metric for performance, but because our target classes are highly imbalanced, I did not do so. Instead, I decided to use F1 as the determining factor of best model performance. For this research, it should be the case that it is similarly as important to have accurate positive predictions (precision) as it is to have complete positive predictions (recall). We want to be sure of hate crimes we associate with a bias and also sure that we have covered all the hate crimes associated with that bias. In other words, the models should avoid classifying racial/ethnic/ancestral biases as religious biases, sexual orientation biases as disability biases, etc. and should also try to identify as many of each bias as possible. To further account for the uneven class distribution, I will be weighting all F1 scores.
   (a) Precision - This metric measures the accuracy of positive predictions made. The calculation for this metric is given by $\frac{\text{TP}}{\text{TP} + \text{FP}}$.
   (b) Recall - This measures the completeness of positive predictions made. The calculation for this metric is given by $\frac{\text{TP}}{\text{TP} + \text{FN}}$.
2. Confusion Matrices - These tables represent the prediction summary of a classification model in matrix form. I used them to help evaluate the performance of my models and also to guide error analysis.

4.4.1. Model Tuning

After initial training and evaluation, I then tuned the hyperparameters of each model, except the

Dummy classifier, using scikit-learn's GridSearchCV[10] to cover ample spaces for each hyperparameter. I selected 5 folds for cross validation and weighted F1 for scoring. To further understand feature imporance, I extracted the top 25 features from my tuned RF model and computed their permutation importances. These are outlined in Section 5.3. Permutation feature importance is defined to be the decrease in a model's score when a single feature value is randomly shuffled.

4.5. Python Packages

The following packages were useful for the implementation of my project:

1. pandas: This was used for data manipulation and analysis.
2. NumPy: This was used for mathematical operations on dataframes.
3. matplotlib: This was used for visualizing data and results.
4. scikit-learn: This was used for creating, training, and tuning most models.
5. imbalanced-learn: This was used for oversampling the imbalanced dataset.
6. keras: This was used for creating, training, and tuning the ANN model.

[10] I automated ANN tuning separately.

5. Results and Evaluation

5.1. Exploratory Data Analysis

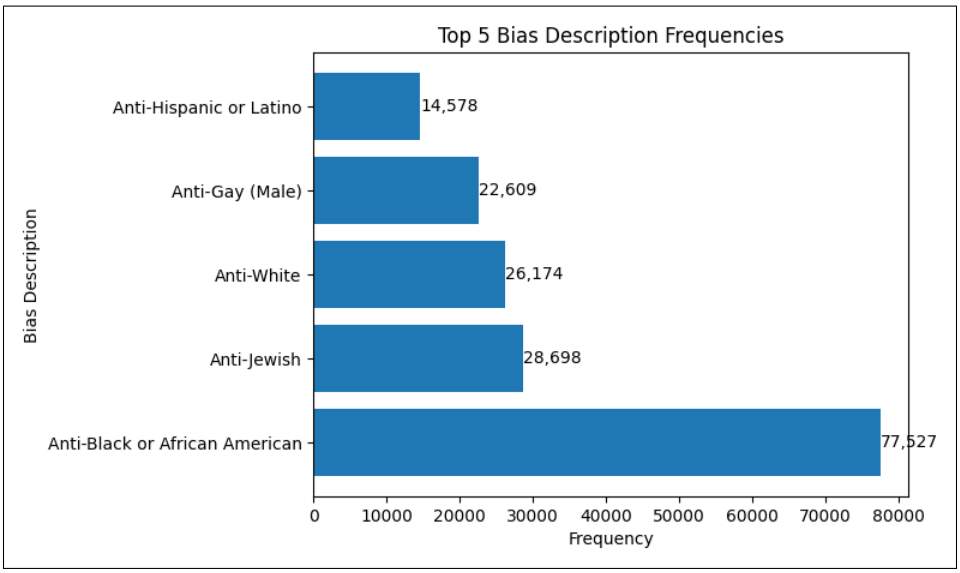
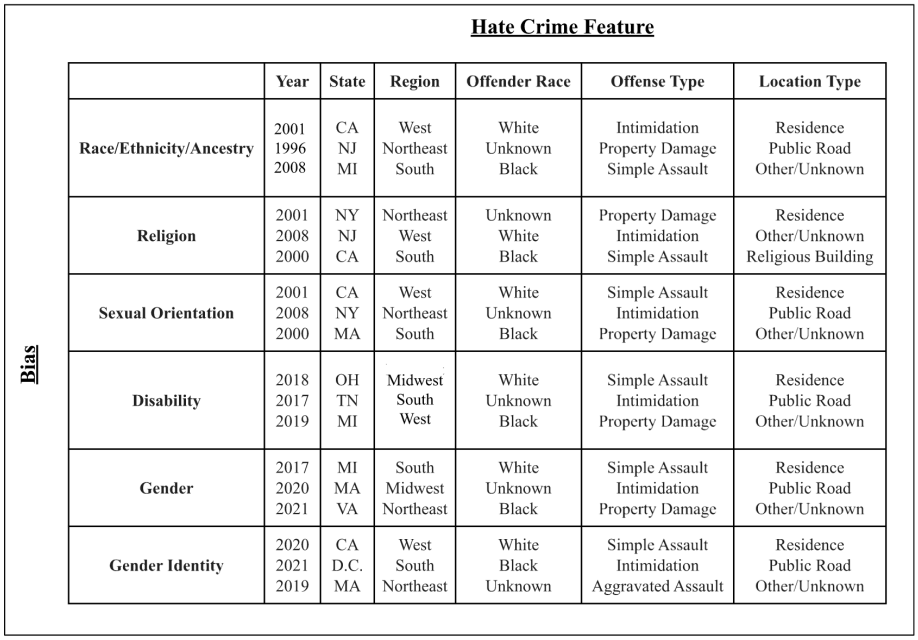


Figure 2: The top 5 hate crime biases.

Examining feature frequencies during exploratory data analysis revealed interesting insights about the dataset. For example, I found that since 1991, hate crimes have occurred the most in the years of 2020 and 2021, the states of California and New Jersey, and the West and Northeast regions. Going even further, hate crimes have most commonly involved offenses of intimidation and destruction of property, locations like the home or public roads, and known offender races of White and African American. It is important to note that for offender race, the "Unknown" category does comprise 40% of the data points which possibly indicates that many offenders are never caught, or hate crimes are reported long after the fact. Perhaps the most important feature I examined, though, was bias description. As shown in Figure 2, approximately a third of the hate crimes in the dataset were motivated by Anti-African American sentiments, followed by large amounts of Anti-Jewish, Anti-White, Anti-Gay (Male), and Anti-Hispanic or Latino hate crimes.



	Year	State	Region	Offender Race	Offense Type	Location Type
Race/Ethnicity/Ancestry	2001 1996 2008	CA NJ MI	West Northeast South	White Unknown Black	Intimidation Property Damage Simple Assault	Residence Public Road Other/Unknown
Religion	2001 2008 2000	NY NJ CA	Northeast West South	Unknown White Black	Property Damage Intimidation Simple Assault	Residence Other/Unknown Religious Building
Sexual Orientation	2001 2008 2000	CA NY MA	West Northeast South	White Unknown Black	Simple Assault Intimidation Property Damage	Residence Public Road Other/Unknown
Disability	2018 2017 2019	OH TN MI	Midwest South West	White Unknown Black	Simple Assault Intimidation Property Damage	Residence Public Road Other/Unknown
Gender	2017 2020 2021	MI MA VA	South Midwest Northeast	White Unknown Black	Simple Assault Intimidation Property Damage	Residence Public Road Other/Unknown
Gender Identity	2020 2021 2019	CA D.C. MA	West South Northeast	White Black Unknown	Simple Assault Intimidation Aggravated Assault	Residence Public Road Other/Unknown

Figure 3: The top 3 most frequent values for hate crime features by bias.

After relabeling the 35 individual biases into 6 groups, I also examined the top 3 most frequent categories of select features for each group as shown in Figure 3 and found interesting trends.[11] Disability, gender, and gender identity biases have been most frequent in the last 6 years compared to the other 3 bias categories which were most frequent in the early 2000s. It is also interesting to see that Anti-Race/Ethnicity/Ancestry biases made a rebound in 2020. As for the region feature, each bias group has been common in the West, Northeast, and South, and it is more interesting to see that the Midwest has had the most hate crimes with disability biases. In terms of offender race, these hate crimes biases are usually held by White and Black people, but many of times, especially in Anti-Religion hate crimes, we see again that the offender's race remains unknown. Finally, across all bias groups, intimidation, property damage, and assault are the common offenses involved, but it is the Anti-Religion bias that most often involves property damage, which makes sense because many religions have buildings for worship.

[11] Note that the Hate Crime Statistics dataset I downloaded does not include complete data from 2021 because the FBI just released the 2021 Supplement.

5.2. Classification Model Performance

5.2.1. Performance Before Tuning

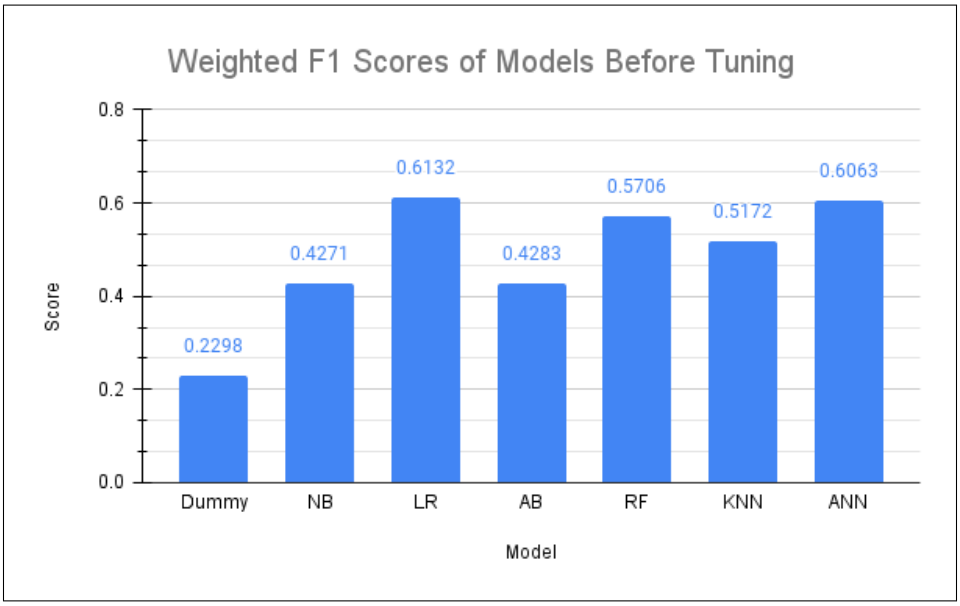


Figure 4: Weighted F1 scores of the models before tuning.

The initial performance of each classifier is shown in Figure 4. As expected, the Dummy classifier had the poorest performance compared to the rest of the models, achieving just over 50% of the next worst performing model (NB)'s weighted F1 score. This difference simply indicates that the features of the dataset do have some predictive power, and we are better off not just randomly predicting. We also observe that of the two linear models NB and LR, LR prevailed by $\sim$ 19% which supports that the lower bias from using LR is beneficial for better predictions. As for the two ensemble classifiers AD and RF, RF prevailed by $\sim$ 14% which may be a result of AD overfitting the data. Since decision trees are iteratively tweaked to focus on difficult areas, AB can provide more accurate predictions on the training data but be more sensitive to overfitting than RF, resulting in a lower performance on the testing set. Overall, LR had the highest weighted F1 score of 61.31%, closely followed by ANN with 60.63%. Considering the high imbalance of classes, these are decent initial results. The initial normalized confusion matrices in Figures 5 and 6 show that NB had a better

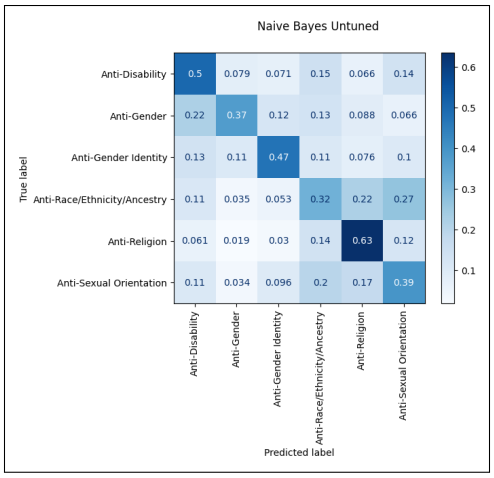


Figure 5: Confusion matrix of untuned NB.

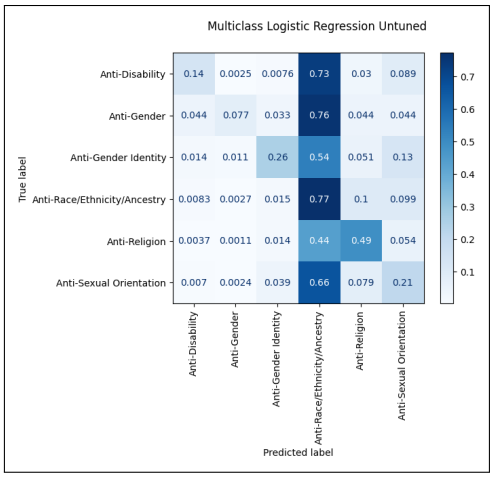


Figure 6: Confusion matrix of untuned LR.

macro recall score than LR, outperforming LR in every class except Anti-Race/Ethnicity/Ancestry. However, LR's weighted recall pulls its weighted F1 score far ahead of NB's because LR had better predictions for Anti-Race/Ethnicity/Ancestry. Similarly, Figures 7 and 8 show that AB had a better

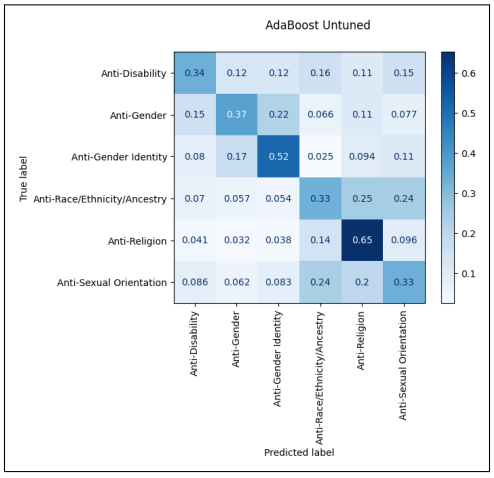


Figure 7: Confusion matrix of untuned AB.

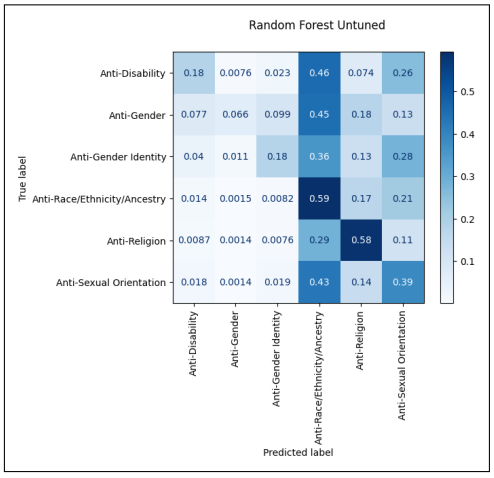


Figure 8: Confusion matrix of untuned RF.

macro recall score than RF, outperforming RF in every class except Anti-Race/Ethnicity/Ancestry and Anti-Sexual Orientation. However, RF's weighted recall score pulls its weighted F1 score far ahead of AB's because RF had better predictions for Anti-Race/Ethnicity/Ancestry.

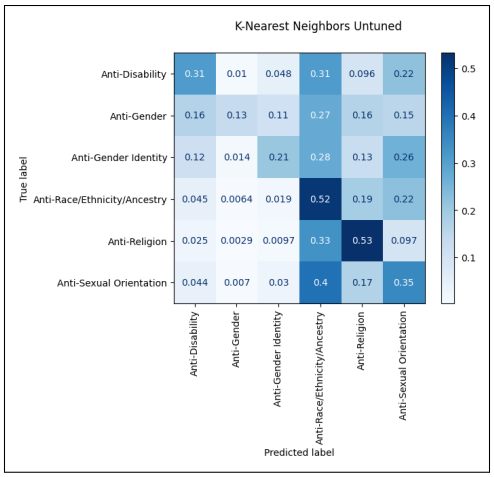


Figure 9: Confusion matrix of untuned KNN.

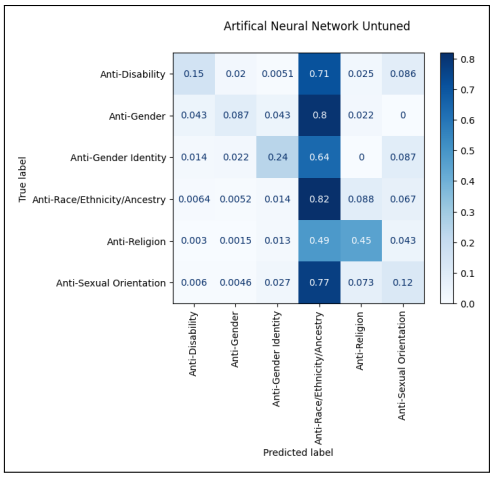


Figure 10: Confusion matrix of untuned ANN.

As for Figures 9 and 10, instead of comparing them to each other, it is interesting to see how similar their confusion matrices look to the previous ones. For example, KNN and RF have similar matrices, while ANN and LR have similar matrices. Additionally, NB and AB have similar matrices.

5.2.2. Performance After Tuning

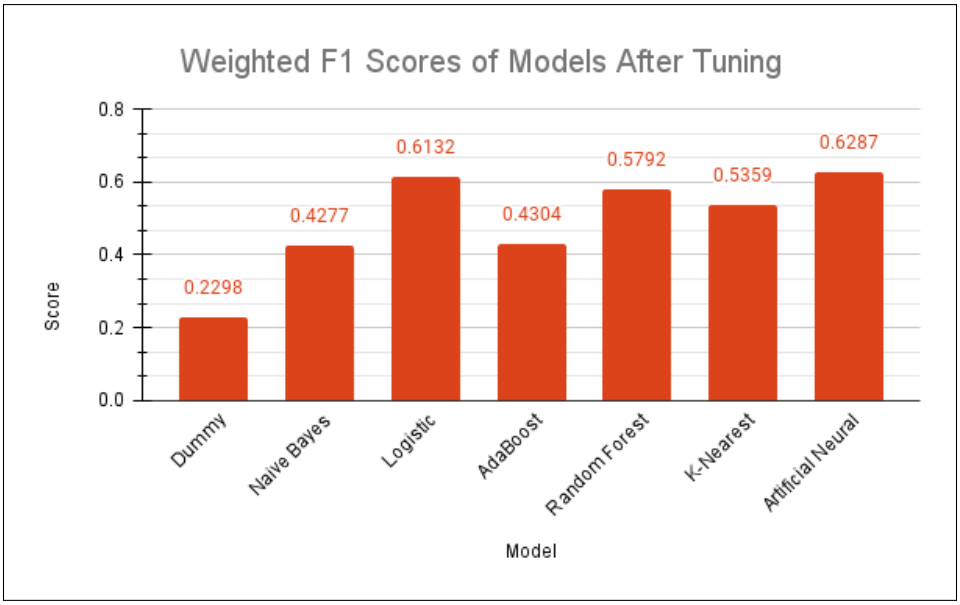


Figure 11: Weighted F1 scores of the models after tuning.

After tuning the hyperparameters of each model, every model except the Dummy classifier and

LR saw improvements in their weighted F1 scores. The relative rankings of each model remained the same except ANN surpassed LR by 1.55%.

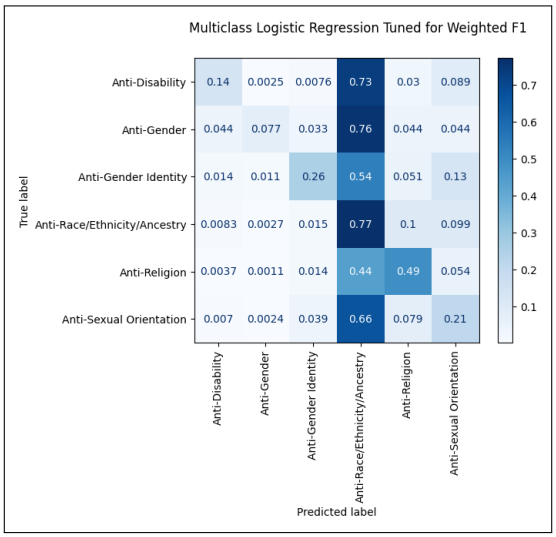


Figure 12: Confusion matrix of tuned LR.

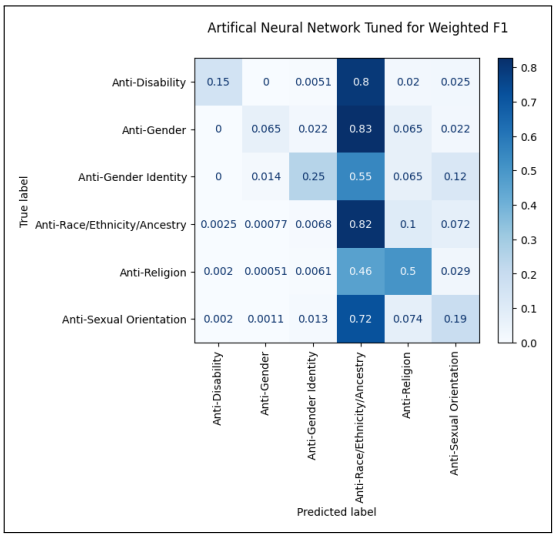


Figure 13: Confusion matrix of tuned ANN.

Since each model's tuned confusion matrix did not change drastically, I only include the matrices of our top 2 models here. The tuned normalized confusion matrices in Figures 12 and 13 show that LR had very similar recall scores to ANN, outperforming ANN in Anti-Gender, Anti-Gender Identity, and Anti-Sexual Orientation by 1-2% margins. However, it was ANN's precision per class that pulled ANN ahead in the final weighted F1 score. Thus, LR and ANN have similar completeness of postiive predictions, but ANN's positive predictions are more accurate when considering class imbalance.

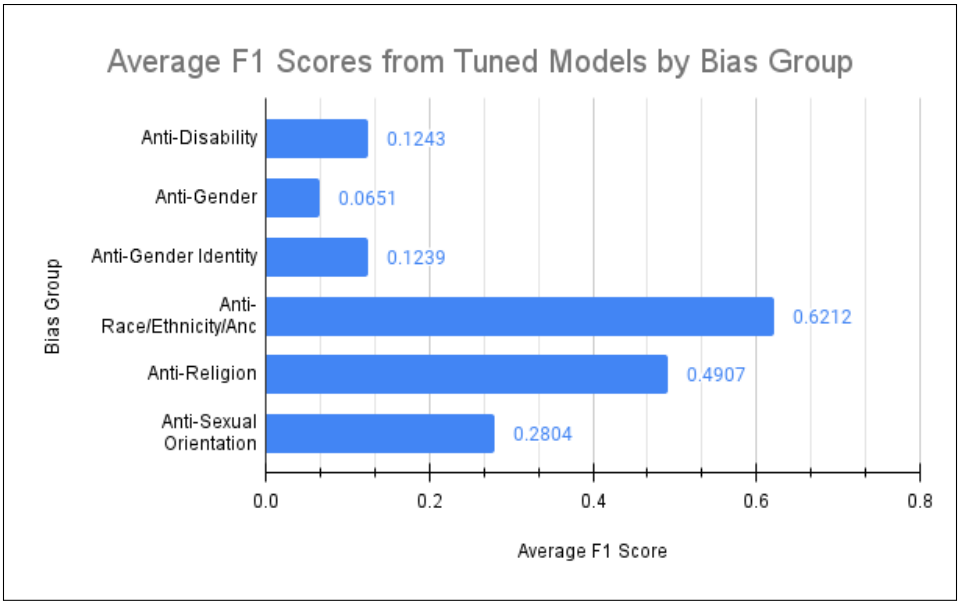


Figure 14: The average F1 scores for each bias group after tuning all models.

I also computed the average F1 score of all tuned models for each bias group, and the rankings align with class proportion. The average F1 score decreases when moving to a class with less samples in the testing set. Thus, it is clear that if we had more samples from each of the non-majority classes, our models could have achieved higher weighted F1 scores.

5.3. Feature Importance

Before training my models, I was able to look at preliminary feature importances using scikit-learn's SelectKBest. The top 25 features were: 2021, District of Columbia, Mississippi, New Jersey, New York, Ohio, Middle Atlantic, Pacific, Midwest, Northeast, West, Cities from 500,000 thru 999,999, Black or African American, Unknown Race, White, All Other Larceny, Destruction/Damage/Vandalism of Property, Simple Assault, Bar/Nightclub, Church/Synagogue/Temple/Mosque, Highway/Road/Alley/Street/Sidewalk, Other/Unknown Location Type, Individual (Victim Type), Other Victim Type, and Religious Organization. Before the model selection and training stage even began, it was apparent that certain geographical locations, offender races, and victim types would be important in discerning different biases in hate crimes.

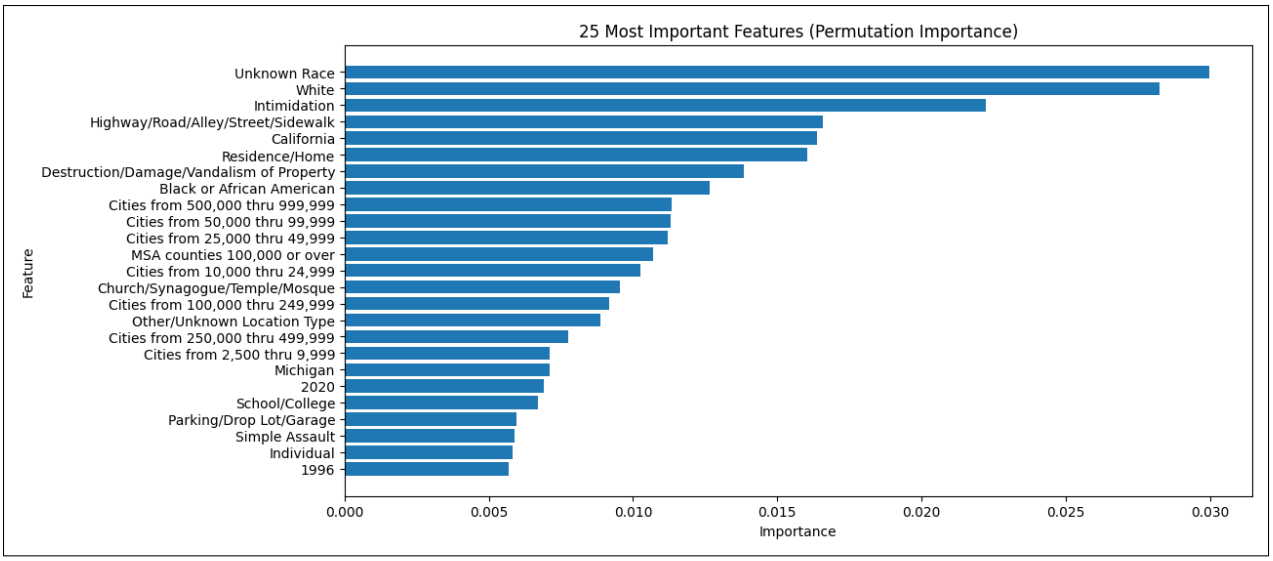


Figure 15: 25 most important features based on permutation importance.

Although my tuned RF model was not the strongest performing classifier, I still found it worthwhile looking at feature importances that I could extract from the model shown in Figure 15. Of the 204 predictive features in the dataset, I found that those most telling of the bias motivating a hate crime related to the offender's race (Unknown, White, Black or African American), the offenses involved (Intimidation, Destruction/Damage/Vandalism of Property, Simple Assault), the location type (Highway/Road/Alley/Street/Sidewalk, Residence/Home, Church/Synagogue/Temple/Mosque), and the surrounding population density (larger cities).

The overlap in the 25 best features from SelectKBest and my RF model were: Unknown Race, White, Highway/Road/Alley/Street/Sidewalk, Destruction/Damage/Vandalism of Property, Black or African American, Cities from 500,000 thru 999,999, Other/Unknown Location Type, Individual (Victim), and Church/Synagogue/Temple/Mosque.

6. Conclusion

Based on my findings, I conclude that hate crime biases show considerable potential to be accurately predicted given the specifics of a hate crime. Even with a heavily imbalanced dataset, my top models LR and ANN achieved weighted F1 scores of approximately 60%. I also conclude that the most important features of the FBI's Hate Crime Statistic dataset in terms of predicting the

bias associated with a hate crime pertain the the offender's race, the type of location where the hate crime was committed, and what type of offense was involved. This can be supported by my exploratory data analysis and feature importance findings.

7. Future Work

7.1. Limitations

Despite the successes of my work, there were definitely some limitations and challenges along the way. Although it is the national repository for hate crime data, the Hate Crime Statistics dataset is not an exhaustive collection of every hate crime that has occurred in the United States. Reporting is voluntary for law enforcement agencies, and the victims themselves may fail to report or recognize that a hate crime occurred. The dataset was also limiting because it contained sparse features like offender ethnicity that I had to remove but I thought would have otherwise been useful in analysis and bias prediction. In a similar vein, I felt that the set of features I was left with after data cleaning was quite small and not the most ideal set given how the models performed. Ideally, I would have wanted the dataset to be as detailed as possible, describing the specific demographics of the victim(s), offender(s), and surrounding area. This would have certainly led to more interesting insights and potential features useful for model prediction. Apart from the dataset itself, I could have tried some combination of undersampling and oversampling to deal with the imbalance to prevent overfitting for the minority classes. Additionally, some of the models I chose were particularly hard to train and tune given time and computational constraints. For example, my Random Forest, K-Nearest Neighbors, and AdaBoost models would often crash my Google Colab sessions after running for several hours. I had to create separate notebooks to train each model, and to tune each one, I had to limit the grid search even if I knew some hyperparameters could be tuned even further to improve performance. In an attempt to increase computational power, I even subscribed to Google Colab Pro, but my runtimes would still disconnect, despite writing code to prevent idleness. I estimate that over 200 hours of training time have been lost along the way.

7.2. Next Steps

Every year, the FBI updates the Hate Crime Statistics dataset, and because of increased efforts to encourage widespread reporting, data is more plentiful every year. To build upon my work in the future, I would like to employ the same methods I used thus far on more updated versions of the dataset once they are released or try using completely different hate crime datasets. Having additional features and data points to work with can potentially strengthen and diversify my classification models and also reveal additional patterns across hate crimes. Another potential avenue I would like to explore is changing the machine learning task to be classifying within each of the six general bias groups focused on in my project. Since each of the target classes is comprised of multiple subclasses, it would be very interesting to research why some races/ethnicities/ancestries, religions, sexual orientations, disabilities, genders, and gender identities are more commonly targeted than others and what the hate crimes associated with them usually entail. There is much work to be done in order to better understand hate crimes, and the possibilities for learning are endless with machine learning.

8. Acknowledgements

There are many people I would like to acknowledge for their valuable contributions towards the success of my project. I first want to express my sincerest gratitude to my adviser Dr. Xiaoyan Li who taught COS IW 03: Machine Learning and Data Science. Every week, she provided our seminar with a plethora of knowledge, resources, and tools to help guide our projects, and she also improved our presentation skills. I also want to thank all of my peers in the seminar for providing me with constructive feedback on my project every week and giving me the opportunity to learn about your own projects. Lastly, I want to thank Meet Patel for being a source of guidance and support throughout the semester as an alumni of the seminar. You have all taught me so much, and I appreciate all of your time and effort spent aiding me in my work.

9. Honor Code

This paper represents my own work in accordance with University regulations. - Jimmy Hoang

References

[1] Federal Bureau of Investigation, "The hate crime statistics dataset," 2021, retrieved from https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/downloads.

[2] S. Han *et al.*, "American hate crime trends prediction with event extraction," *arXiv preprint arXiv:2111.04951*, 2021.

[3] M. Jendryke and S. C. McClure, "Spatial prediction of sparse events using a discrete global grid system; a case study of hate crimes in the usa," *International Journal of Digital Earth*, vol. 14, no. 6, pp. 789–805, 2021.

[4] S. Wang, "Hate crime analysis based on artificial intelligence methods," in *E3S Web of Conferences*, vol. 251. EDP Sciences, 2021, p. 01062.

10. Appendix

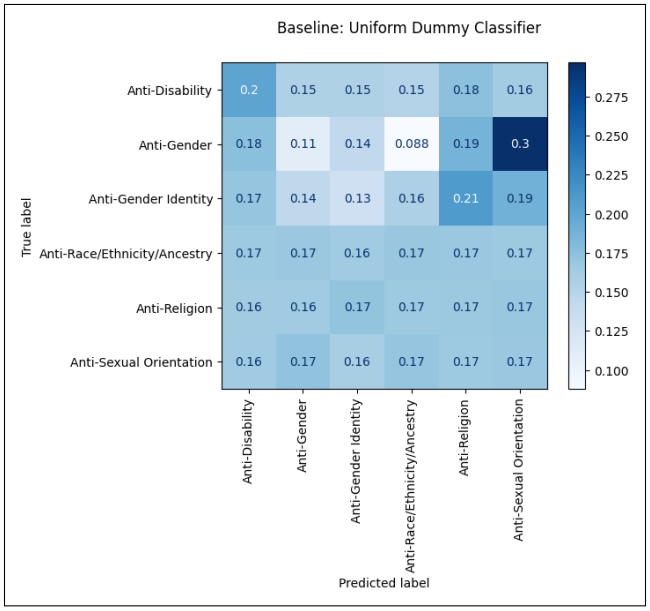


Figure 16: Confusion matrix of the Dummy classifier.

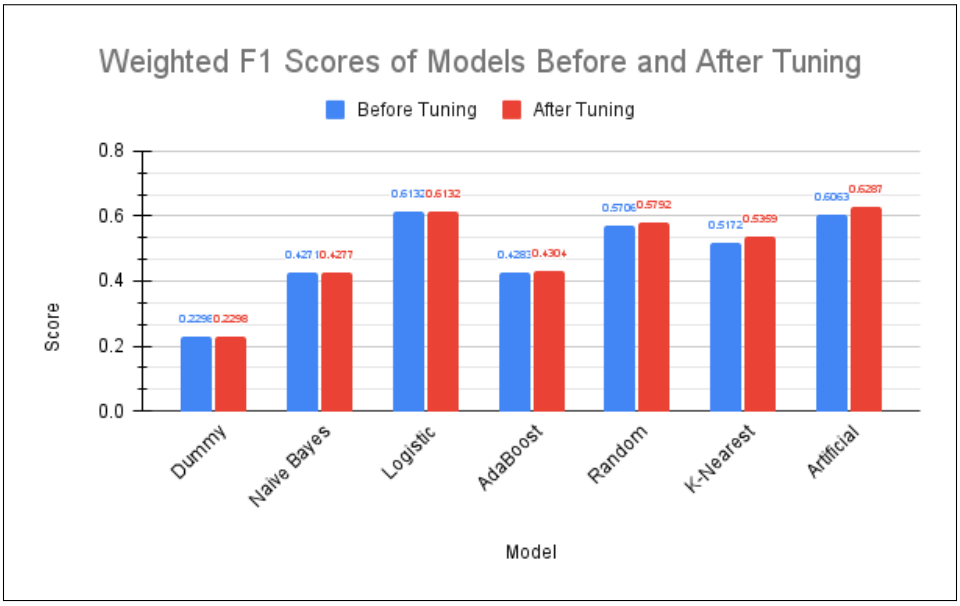


Figure 17: Weighted F1 scores of models before and after tuning.

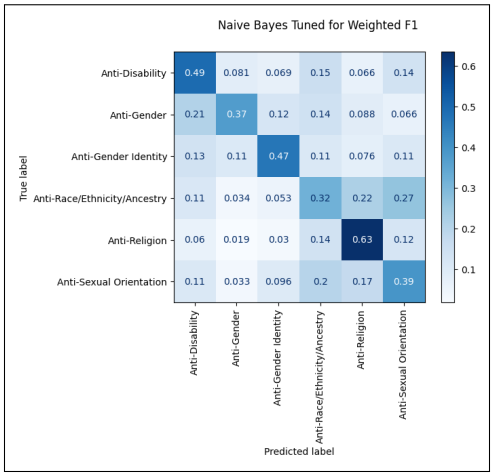


Figure 18: Confusion matrix of tuned NB.

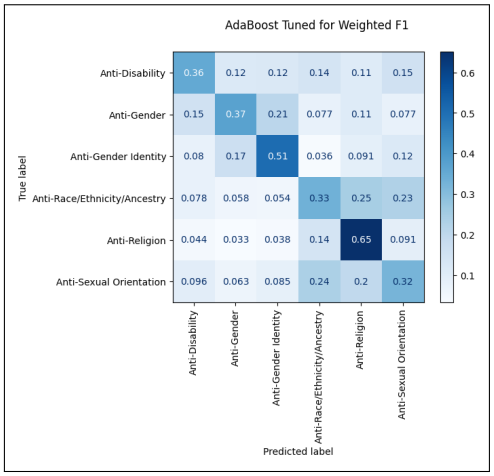


Figure 19: Confusion matrix of tuned AB.

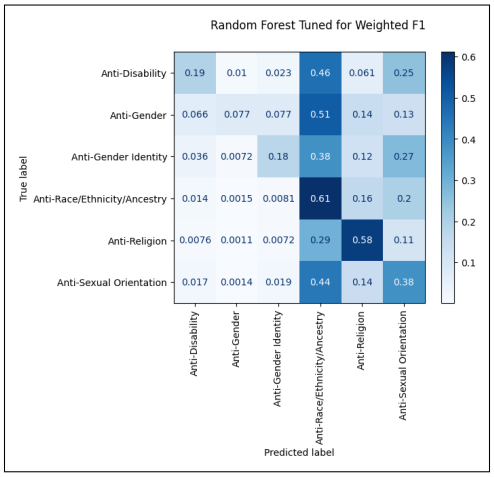


Figure 20: Confusion matrix of tuned RF.

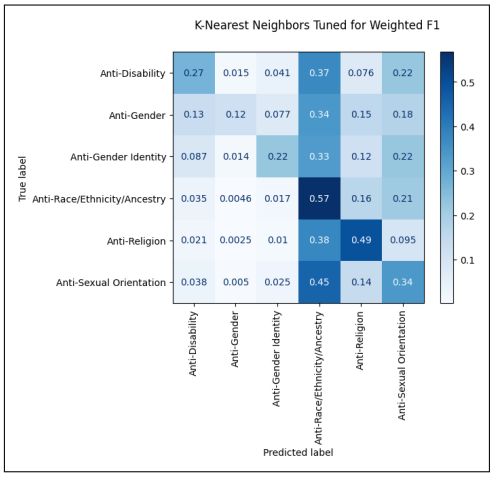


Figure 21: Confusion matrix of tuned KNN.