Project: Cyberbullying Detection and its Type

Abstract

This project is based on the development of a data science project, which uses Machine Learning, and NLP tools and techniques to detect cyberbullying. In this project, we have two different labeled datasets. One is used for identifying whether a text has cyberbullying in it and the second one identifies the specific type of cyberbullying, such as race, religion, gender and age. Dataset is passed through different phases. Firstly, data preprocessing techniques are applied to it, including lowercasing, regular expression-based cleaning, expansion of contractions and slangs, and handling of munched words. Various attributes are extracted from it, including density of offensive words, sentimental analysis, and weighted average of offensive words. Different classification models are used to predict cyberbullying and the specific type, including SVC, Logistic Regression, Random Forest Classifier, KNN, Naive Bayes, and Extra Trees Classifier. Results show that Random Forest Classifier achieved the highest accuracy of 85% in identifying cyberbullying and logistic regression achieved 94% accuracy for classifying the type of cyberbullying.

Purpose

In order to make online spaces more hospitable, we endeavor to construct a project which leverages machine learning and natural language processing to successfully spot and prevent cyberbullying. The plan involves identifying malicious content with precision and categorizing the kind of cyberbullying identified, ultimately forming a more secure and welcoming digital atmosphere. Our main objective is to nurture users' mental health and overall wellness by preventing instances of bullying, encouraging common living, and providing tools for positive and respectful interactions within the system.

Our main purpose is to create a social media platform that detects and handle cyberbullying. NLP allows us to interpret and extract meaning from textual data, and evaluate linguistic structure and text pattern. Machine learning helps us predict unseen data. With the help of machine learning models and NLP techniques, we can address online abuse and harm and get rid of it for good.

Objectives

- The main idea is to create a project that accurately detects cyberbullying and determine the type of it present in the text.
- By detecting the type of bullying, including race, religion, age, and gender, we can create tailored responses, provide effective moderation, and can gain insights about a specific type.
- Use Natural Language Processing (NLP) to analyze the textual data and detect cyberbullying from it through linguistic patterns.
- Use Machine learning models can be used to predict whether a text can be deemed as offensive or safe.
- Create a positive project by removing online hate, and abuse, so users are free from any potential harm and abuse.
- Continuously improve and evolve our project to keep up with new patterns of abusive language and to provide the best possible service to our clients.

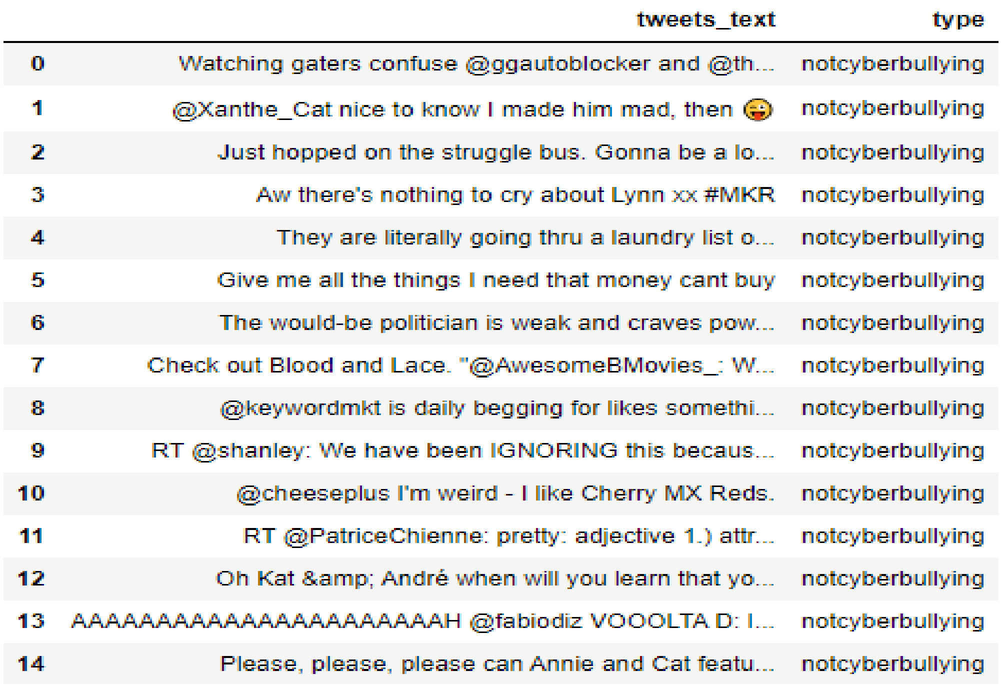
Libraries and Technologies

- Python
- Jupyter Notebook
- MySQL
- Re
- Nltk
- Vader
- Pandas
- Scikit-learn
- Pickle
- WordNinja

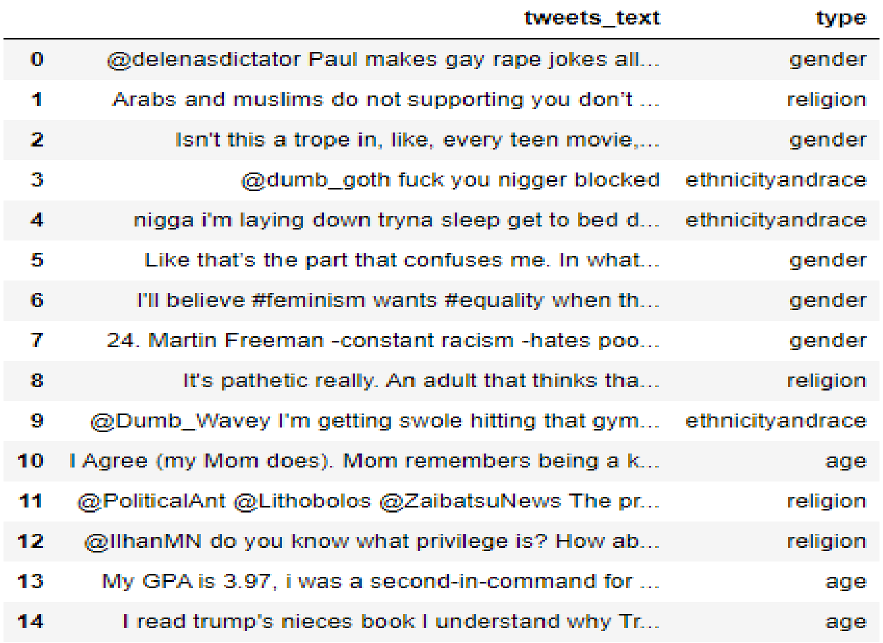
Steps Documented

1) Data Collection:

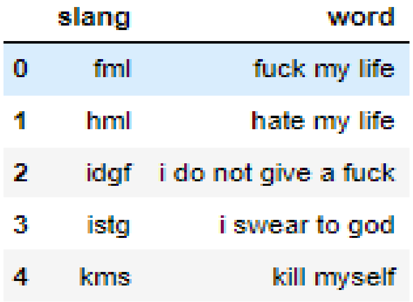
Data collected for this project was taken from Kaggle. It is a dataset which contains data, collected from twitter. The initial dataset contained more than 47000 tweets, which classified tweets according to the class of cyberbullying; age, ethnicityandrace, gender, religion, others and not cyberbullying. Due to the abundant amount of data, I decided to divide the dataset into two parts; first one will be used to decide whether cyberbullying exists and if it does, the second one will be used to determine the type of cyberbullying available. I decided to keep roughly 250 instances of each class, rather than going forward with using all 47000 instances. I have collected different wordlists of age, race, gender and religion glossary. Also, a list of offensive words with their respective severity, a slang abbreviated list, and a negation list. All of the data is stored in the database in form of different tables.



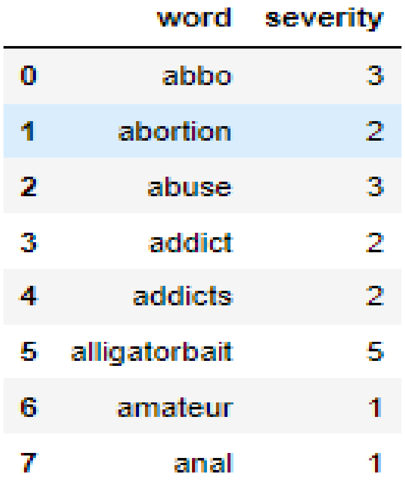
	tweets_text	type
0	Watching gaters confuse @ggautoblocker and @th...	notcyberbullying
1	@Xanthe_Cat nice to know I made him mad, then 😜	notcyberbullying
2	Just hopped on the struggle bus. Gonna be a lo...	notcyberbullying
3	Aw there's nothing to cry about Lynn xx #MKR	notcyberbullying
4	They are literally going thru a laundry list o...	notcyberbullying
5	Give me all the things I need that money cant buy	notcyberbullying
6	The would-be politician is weak and craves pow...	notcyberbullying
7	Check out Blood and Lace. "@AwesomeBMovies_: W...	notcyberbullying
8	@keywordmkt is daily begging for likes somethi...	notcyberbullying
9	RT @shanley: We have been IGNORING this becaus...	notcyberbullying
10	@cheeseplus I'm weird - I like Cherry MX Reds.	notcyberbullying
11	RT @PatriceChienne: pretty: adjective 1.) attr...	notcyberbullying
12	Oh Kat &amp; André when will you learn that yo...	notcyberbullying
13	AAAAAAAAAAAAAAAAAAAAAAH @fabiodiz VOOOLTA D: I...	notcyberbullying
14	Please, please, please can Annie and Cat featu...	notcyberbullying



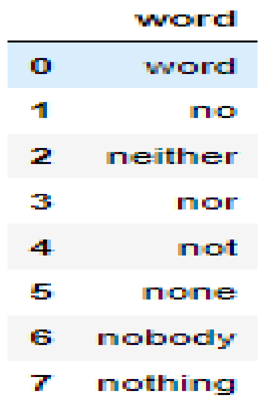
	tweets_text	type
0	@delenasdictator Paul makes gay rape jokes all...	gender
1	Arabs and muslims do not supporting you don't ...	religion
2	Isn't this a trope in, like, every teen movie,...	gender
3	@dumb_goth fuck you nigger blocked	ethnicityandrace
4	nigga i'm laying down tryna sleep get to bed d...	ethnicityandrace
5	Like that's the part that confuses me. In what...	gender
6	I'll believe #feminism wants #equality when th...	gender
7	24. Martin Freeman -constant racism -hates poo...	gender
8	It's pathetic really. An adult that thinks tha...	religion
9	@Dumb_Wavey I'm getting swole hitting that gym...	ethnicityandrace
10	I Agree (my Mom does). Mom remembers being a k...	age
11	@PoliticalAnt @Lithobolos @ZaibatsuNews The pr...	religion
12	@IlhanMN do you know what privilege is? How ab...	religion
13	My GPA is 3.97, i was a second-in-command for ...	age
14	I read trump's nieces book I understand why Tr...	age



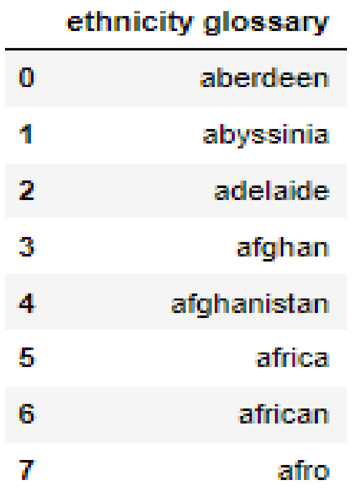
	slang	word
0	fml	fuck my life
1	hml	hate my life
2	idgf	i do not give a fuck
3	istg	i swear to god
4	kms	kill myself



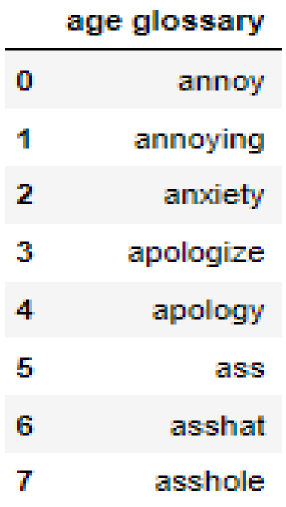
	word	severity
0	abbo	3
1	abortion	2
2	abuse	3
3	addict	2
4	addicts	2
5	alligatorbait	5
6	amateur	1
7	anal	1



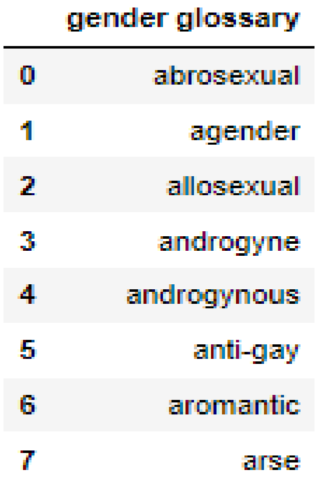
	word
0	word
1	no
2	neither
3	nor
4	not
5	none
6	nobody
7	nothing



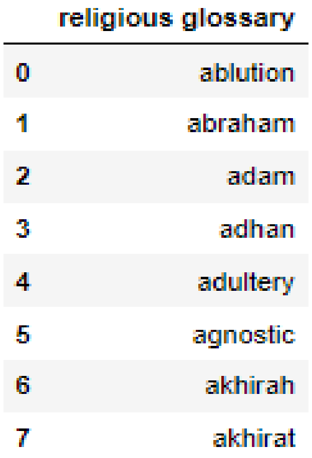
	ethnicity glossary
0	aberdeen
1	abyssinia
2	adelaide
3	afghan
4	afghanistan
5	africa
6	african
7	afro



	age glossary
0	annoy
1	annoying
2	anxiety
3	apologize
4	apology
5	ass
6	asshat
7	asshole



	gender glossary
0	abrosexual
1	agender
2	allosexual
3	androgyne
4	androgynous
5	anti-gay
6	aromantic
7	arse



	religious glossary
0	ablution
1	abraham
2	adam
3	adhan
4	adultery
5	agnostic
6	akhirah
7	akhirat

2) Data Preprocessing:

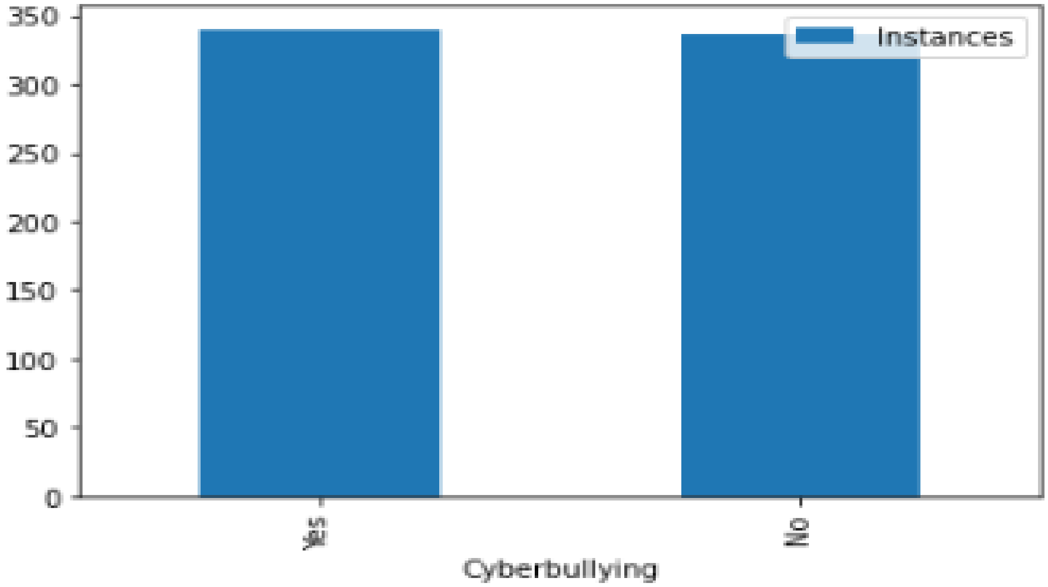
We'll start our data preprocessing by converting our dataset into **lowercase** because we want to scan the words against multiple wordlists. We will then use regular expressions to **clean** the data. As it is obtained through twitter, we need to get rid of @, digits, RT, links or any special characters. We will expand **contractions** (aren't -> are not), so we can get accurate amount of words in a single instance. Also, we will expand **slangs** (hml -> hate my life), so we can detect offensive words from abbreviated slangs. We will then use wordninja library to expand **munched** words (hatemylife -> hate my life), so we can identify slang from munched words. After that we will perform tokenization to create tokens and move to the next step.

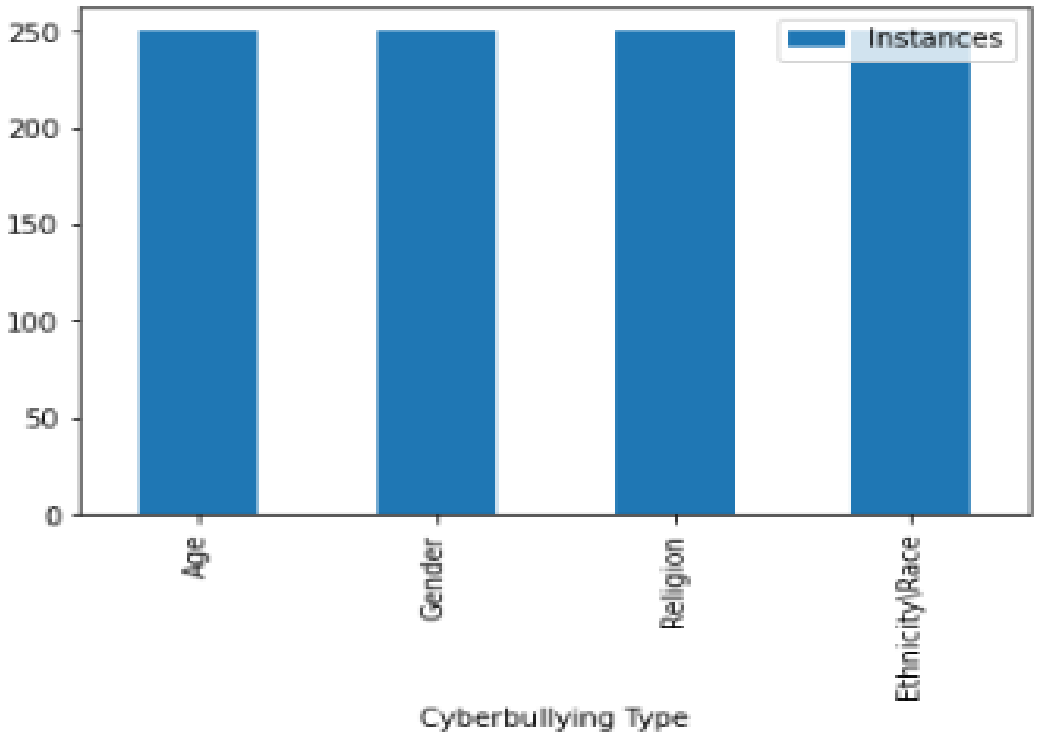
3) Data Transformation:

To feed data to our model, it needs to be in form of numbers, rather than words. So we will create numerical data. Our features include total words and count of offensive words, from which we will create our first feature density of offensive words. We will then calculate the polarity of words by performing sentimental analysis, which will be our second feature. We will then calculate severity of offensive words, by calculating weighted average of them, which will be our third and final feature to determine cyberbullying or not. If our model determines whether a text is deemed offensive, then it will extract these 4 new features to determine the type of bullying present. Four new features include age, race, religion and ethnicity word count, which is just words found in the text and the glossary.

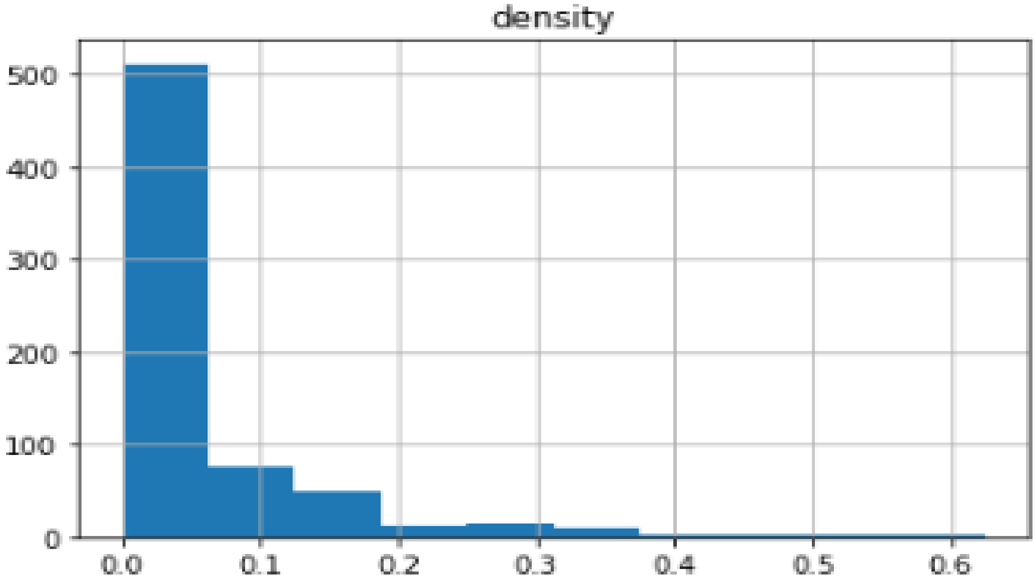
4) Data Visualization:

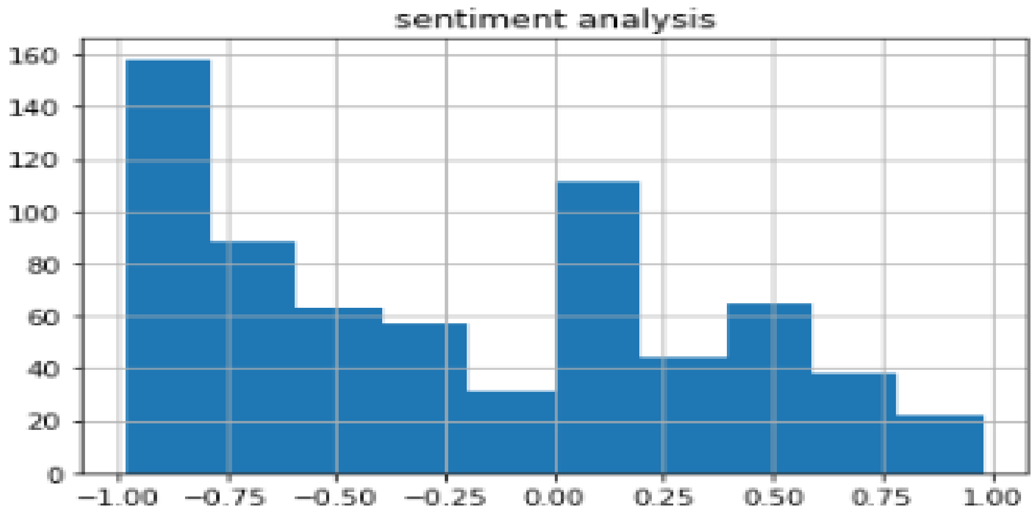
Data visualization before our numerical features:

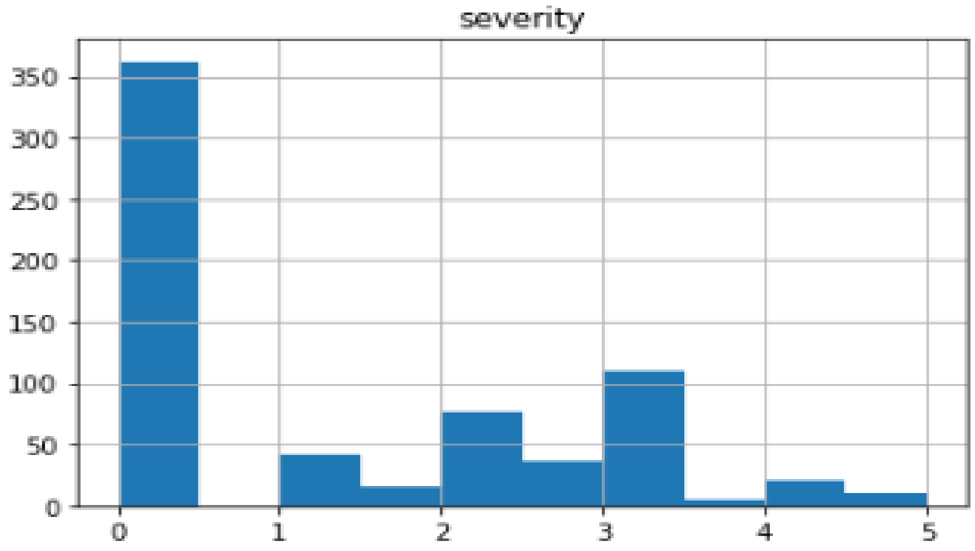


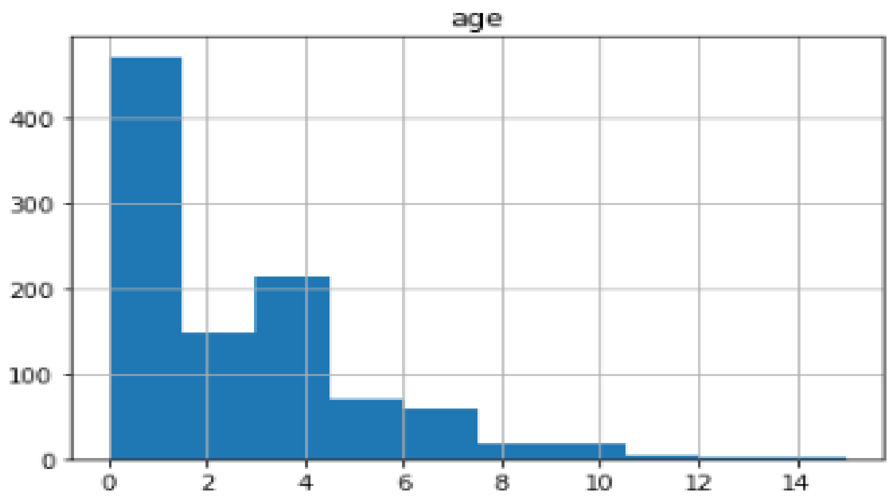


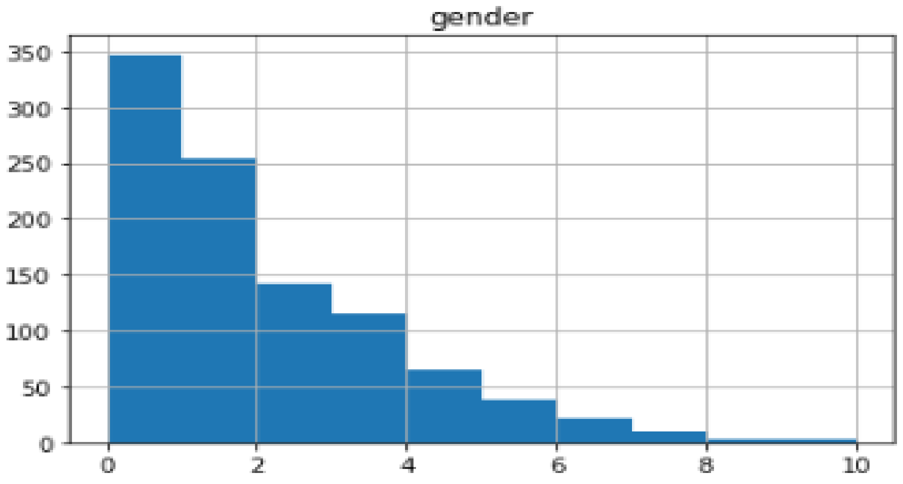
Data visualization after our numerical features:

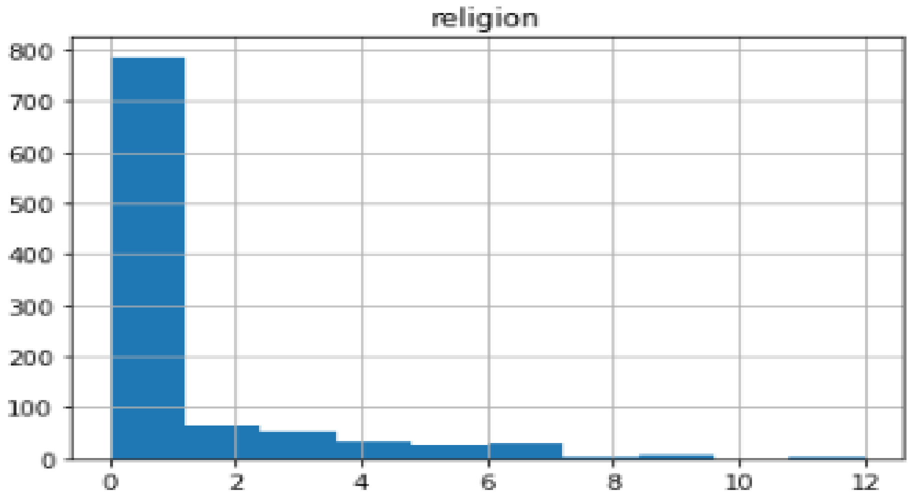


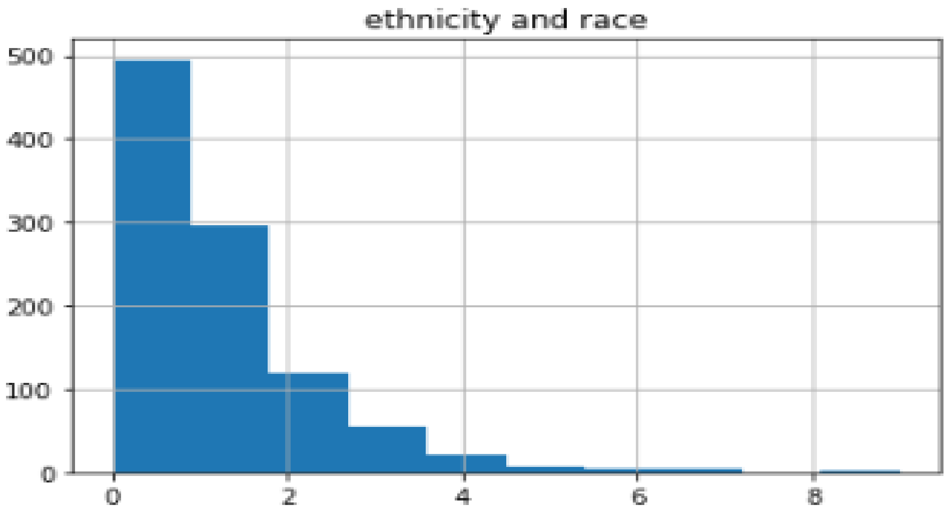

sentiment analysis
160
140
120
100
80
60
40
20
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


severity
350
300
250
200
150
100
50
0
0
1
2
3
4
5


age
400
300
200
100
0
0
2
4
6
8
10
12
14


gender
0
50
100
150
200
250
300
350
0
2
4
6
8
10


religion
0
100
200
300
400
500
600
700
800
0
2
4
6
8
10
12

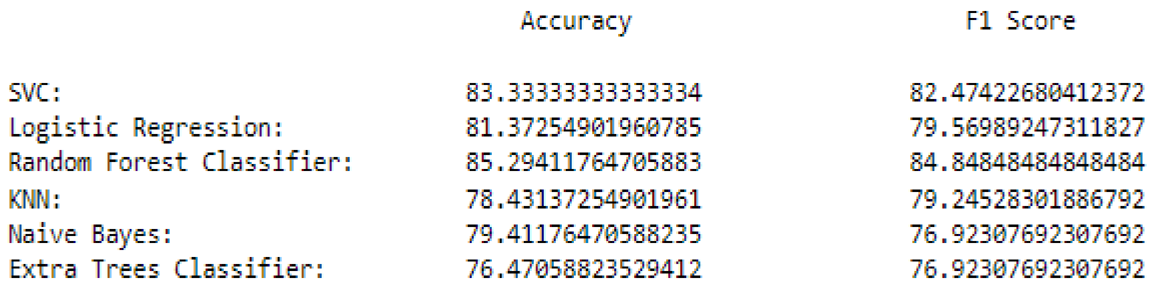

ethnicity and race
0
100
200
300
400
500
0
2
4
6
8

5) Model Development:

We will then create our model to perform predictions by creating our training and testing data to train and evaluate our model. We will split our dataset by train-test split dividing it into 85% train and 15% split while making sure that the split is balanced. We will then use 6 different models to compare scores; SVC, Logistic Regression, Random Forest Classification, KNN, Naïve Bayes and Extra Tree Classifier.

6) Model Evaluation:

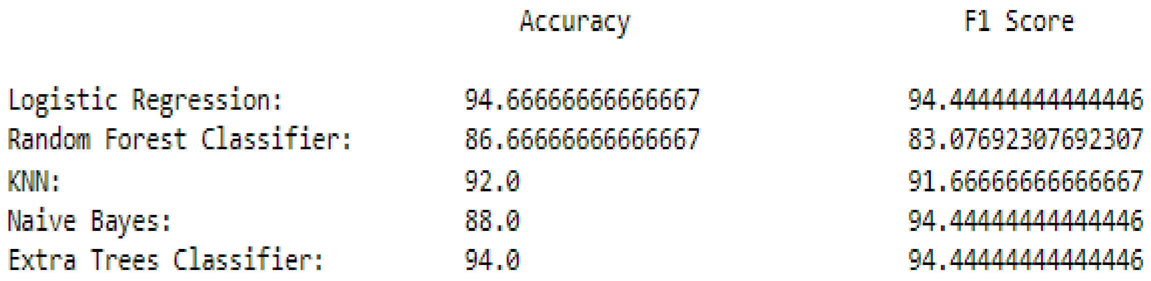
I have picked accuracy and F1 score as the evaluation metrics for these models. Accuracy shows the proportion of correctly classified instances, while F1 score balances precision and recall to assess the performance of a classification model. We tested 6 different models to compare scores; SVC, Logistic Regression, Random Forest Classification, KNN, Naïve Bayes and Extra Tree Classifier. For our cyberbullying classification model, Random Forest Classifier gave the best score.



```
                             Accuracy                 F1 Score

SVC:                         83.33333333333334        82.47422680412372
Logistic Regression:         81.37254901960785        79.56989247311827
Random Forest Classifier:    85.29411764705883        84.84848484848484
KNN:                         78.43137254901961        79.24528301886792
Naive Bayes:                 79.41176470588235        76.92307692307692
Extra Trees Classifier:      76.47058823529412        76.92307692307692
```

For our type classification model, Logistic Regression provided the best scores.



```
                             Accuracy                 F1 Score

Logistic Regression:         94.66666666666667        94.44444444444446
Random Forest Classifier:    86.66666666666667        83.07692307692307
KNN:                         92.0                     91.66666666666667
Naive Bayes:                 88.0                     94.44444444444446
Extra Trees Classifier:      94.0                     94.44444444444446
```

Conclusion

With the help of data preprocessing, feature extraction, and classification model, we were able to create a data science project that identifies and potentially eliminate cyberbullying. We used Python, Jupyter Notebook, and various python libraries to handle our textual data and convert it into suitable format for our models to implement. Our models worked great in detecting and determining the type of cyberbullying present with exceptional accuracy. Our main mission is to create a positive and safe environment for our users, and to eliminate negativity and toxicity of cyberbullying forever.