Hate Speech detection using Transformers (Deep Learning)

Your Deep Learning Partner

Richard Flores & Christos Christoforou

Team Speechium

Christos Christoforou

Richard Flores

Data Glacier

Your Deep Learning Partner

Christos Christoforou

- MSc Computational Applied Mathematics at University of Edinburgh
- BSc Mathematics and Statistics at University of Cyprus
- AI Resident @ Apziva
- Data Science Intern @ Data Glacier
- Interests: Mathematics, Statistics, Machine Learning, AI, NLP, Computer Vision

Data Glacier
Your Deep Learning Partner

Richard Flores

- PhD Data Science Student at National University
- MSc Data Analytics from Western Governors University
- Data Analyst @ Twitter
- Research Assistant @ University of Texas of El Paso
- Interests: Machine Learning, NLP, AI Prompt Engineering, Blockchain Secure Contract Development

Data Glacier
Your Deep Learning Partner

Outline

- Problem description
- Data understanding
- Data cleansing and transformation
- Model Building & Training
- Model Evaluation & Selection
- Model Deployment

Problem description

- Hate speech is any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are.

- In this problem, we will take you through a hate speech detection model with Machine Learning and Python.

- Hate Speech Detection is generally a task of sentiment classification. So, for the task of hate speech detection model, we will use the Twitter tweets to identify tweets containing Hate speech.

Data understanding

- Features:
  - `id’: the primary key,
  - `label’: 0 for free speech and 1 for hate speech,
  - `tweet’: the tweet we want to classify.

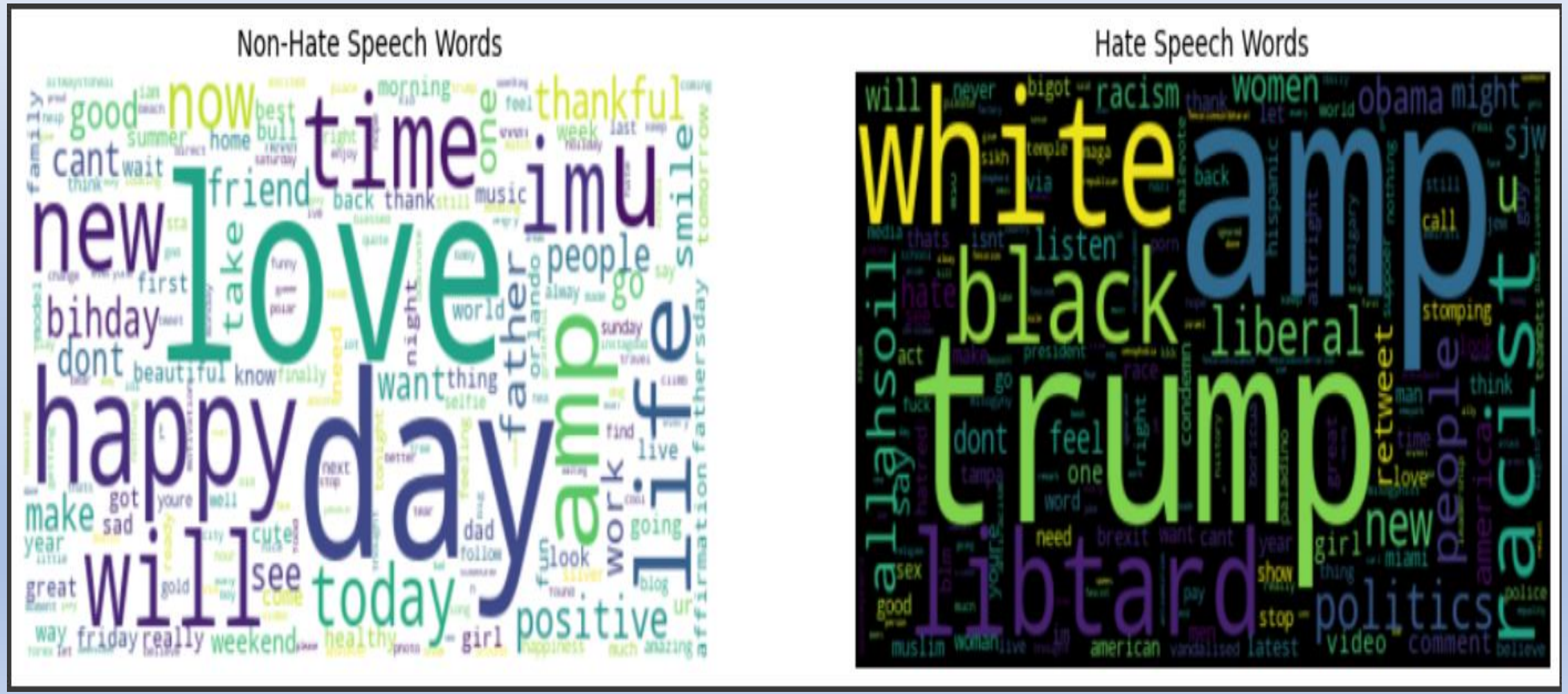
- Feature types:
  - `id’: int64,
  - `label’: int64,
  - `tweet’ : object.

- Null values: There were 0 Null values in the data.

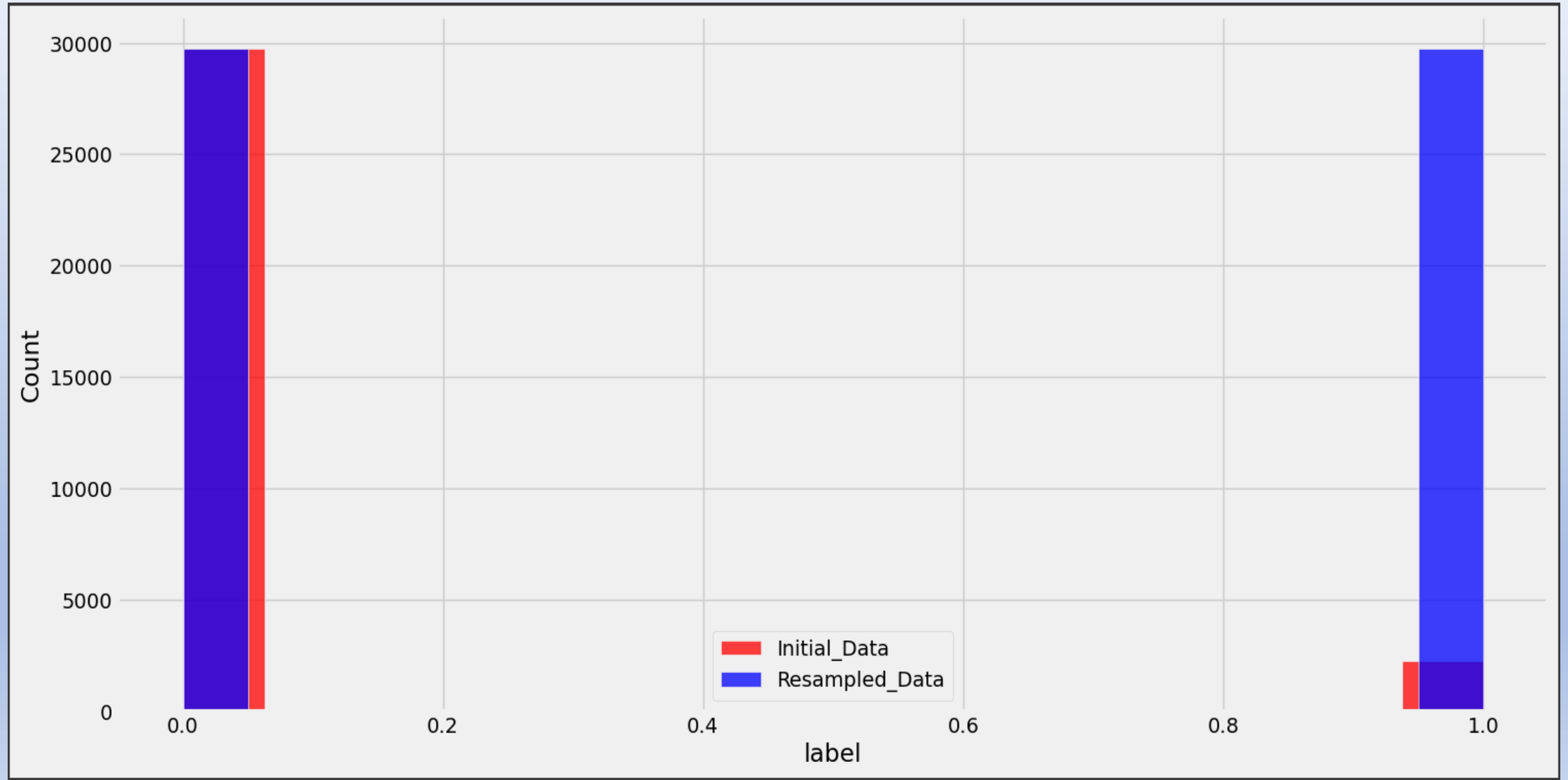
- Duplicated rows: 2432 duplicated rows were found in the data. There are 31962 rows in total.

- Imbalanced data:
  - 27517 tweets belong to the free speech class,
  - 2013 tweets belong to the hate speech class.

A visualization of Non-Hate Speech Words and Hate Speech Words using a Word Cloud.



A visualization of the initial unbalanced data and the resampled balanced data.



Data cleansing and transformation

- **Standard Data Cleaning and Transformation:**
  - Verify data types,
  - Remove NULL values (if any),
  - Remove duplicated data (if any),
  - Resample the data to balance them.

- **NLP Specific Data Cleaning and Transformation:**
  - Apply Tokenization and Lemmatization,
  - Lowercase the text,
  - Remove stop-words and one-letter words,
  - Remove tags and other special characters,
  - Remove non-ASCII characters,
  - Vectorize the training data using the TfidfVectorizer.

Model Building & Training

We built and trained the following models:

- XGBoost Model
  - 5-fold repeated cross-validation

- Simple Transformer Model
  - Adam optimizer
  - Cross-Entropy loss function
  - 5 epochs

- Pretrained Roberta-base Model
  - Adam optimizer
  - Cross-Entropy loss function
  - 4 epochs

Model Evaluation

We will use a range of scoring metrics to evaluate our models. Some of them are: Precision, Recall, and F1-score.

Here are the formulas for each one:

- recall = TP / (TP + FN)
- precision = TP / (TP + FP)
- F1-score = 2 * ((precision * recall) / (precision + recall)), where

TP, TN, FP, and FN are the True Positives, True Negatives, False Positives, and False Negatives respectively. The table below, also known as a confusion matrix, explains each one of them.

True Class	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

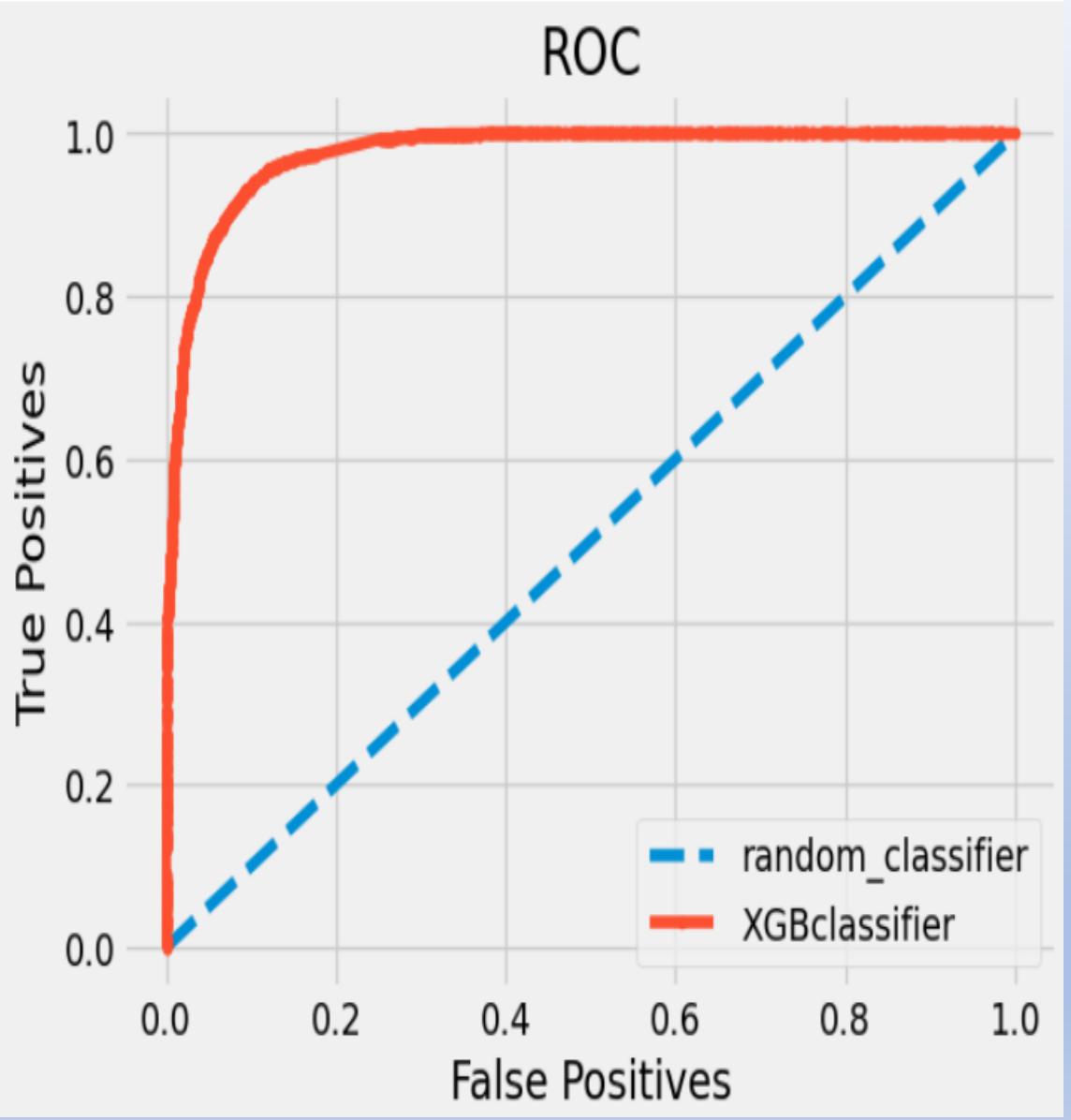
XGBoost Model Evaluation

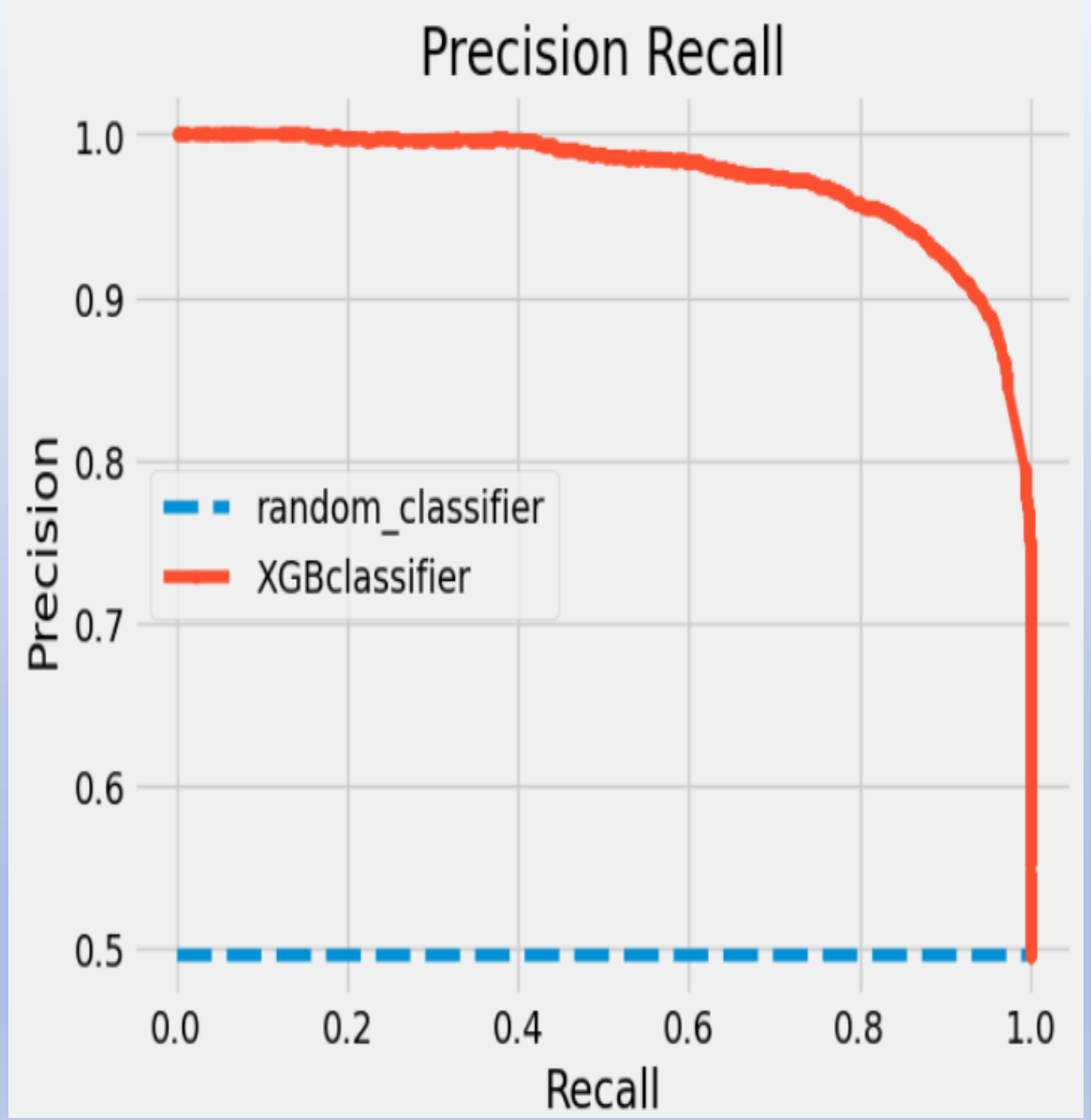
Confusion Matrix:

TP **6787**	**FN** **662**
FP 583	TN 6828

- Precision score: 0.91
- Recall score: 0.92
- AUC score: 0.97
- F1 score: 0.92

```
random_classifier: ROC AUC=0.500
XGBclassifier: ROC AUC=0.976
```





Simple Transformer Model Evaluation

	precision	recall	f1-score
0	0.67	0.47	0.56
1	0.29	0.49	0.37
accuracy			0.48

Roberta-base Model Evaluation

Recall: 0.91	Precision: 0.89
Accuracy: 0.9	F1-score: 0.9
AUROC: 0.95	AUPRC: 0.96

Model Selection

- The XGBoost Model performance was the best of all considered models.

- Hence, we will use the XGBoost Model to predict on unseen data.

Conclusion

- To detect hate speech in tweets, businesses may use a combination of automated tools and human moderation. Automated tools may include machine learning algorithms that are trained to identify hate speech based on certain characteristics, such as the use of certain words or phrases. Human moderation may involve a team of moderators who review tweets and take appropriate action, such as deleting the tweet or banning the user.

- It's important to note that detecting hate speech can be challenging, as it may involve complex issues of context and intent. It is also important for businesses to consider the potential for false positives and ensure that their approaches to detecting and addressing hate speech are fair and transparent.

- In our case, for the data set we have, it seems that simpler models perform better than more complicated ones. Thus, we recommend using a model such as the XGBoost model to predict on this data set.

Thank You

Data Glacier
Your Deep Learning Partner