

Hate Speech detection using Transformers (Deep Learning)



Data Glacier

Your Deep Learning Partner

Richard Flores & Christos Christoforou

Team Speechium

Christos Christoforou



Richard Flores



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



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



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.

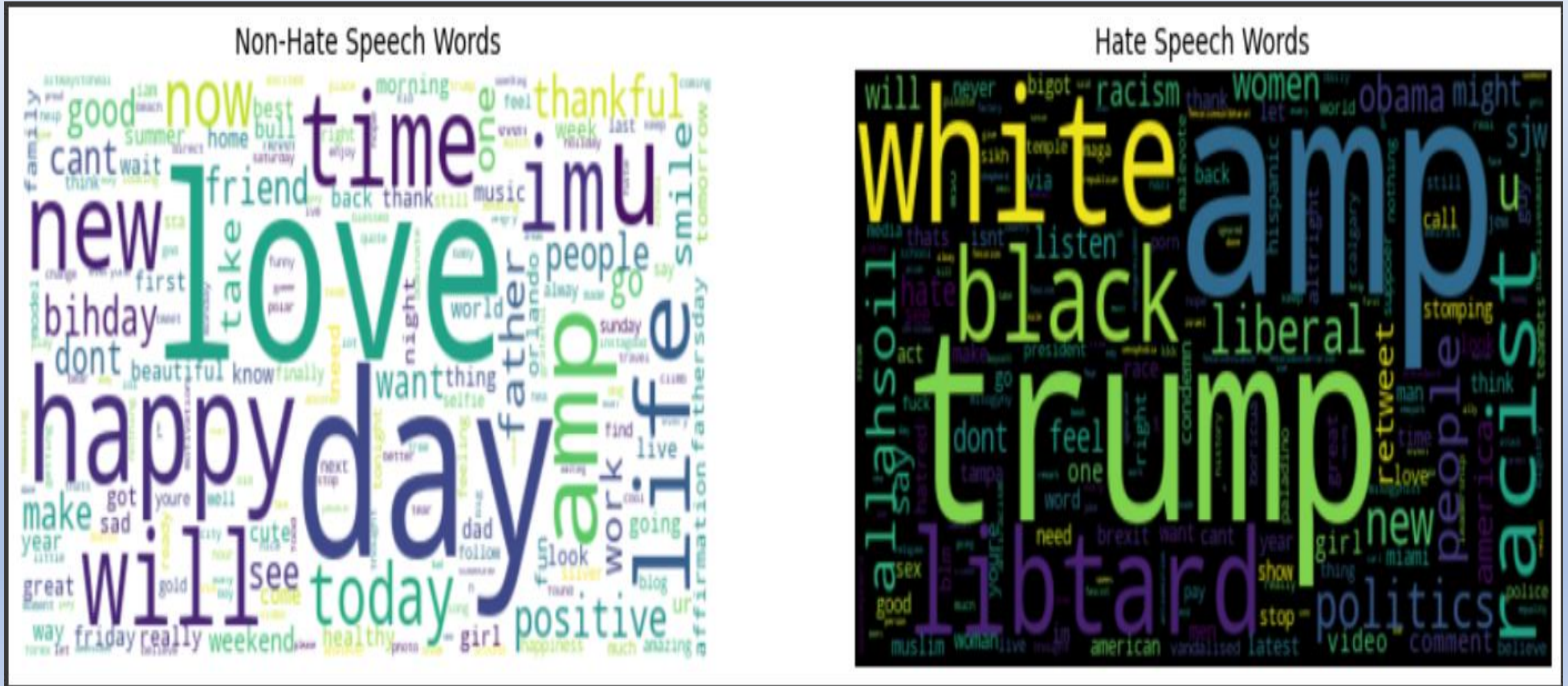
❑ Null values: There were 0 Null values in the data.

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

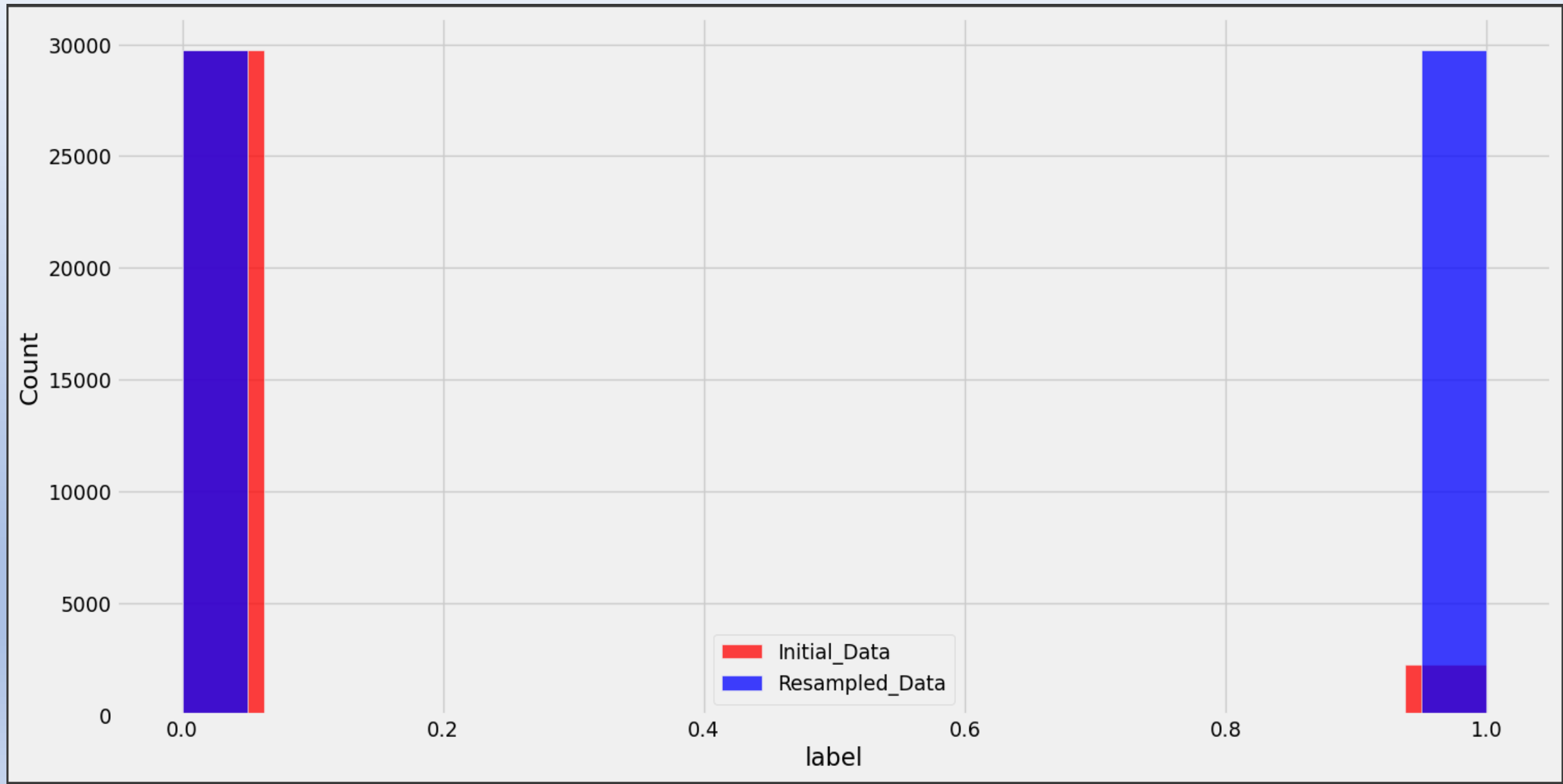
❑ Imbalanced data:

- 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:

- $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$
- $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$
- $\text{F1-score} = 2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$, 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

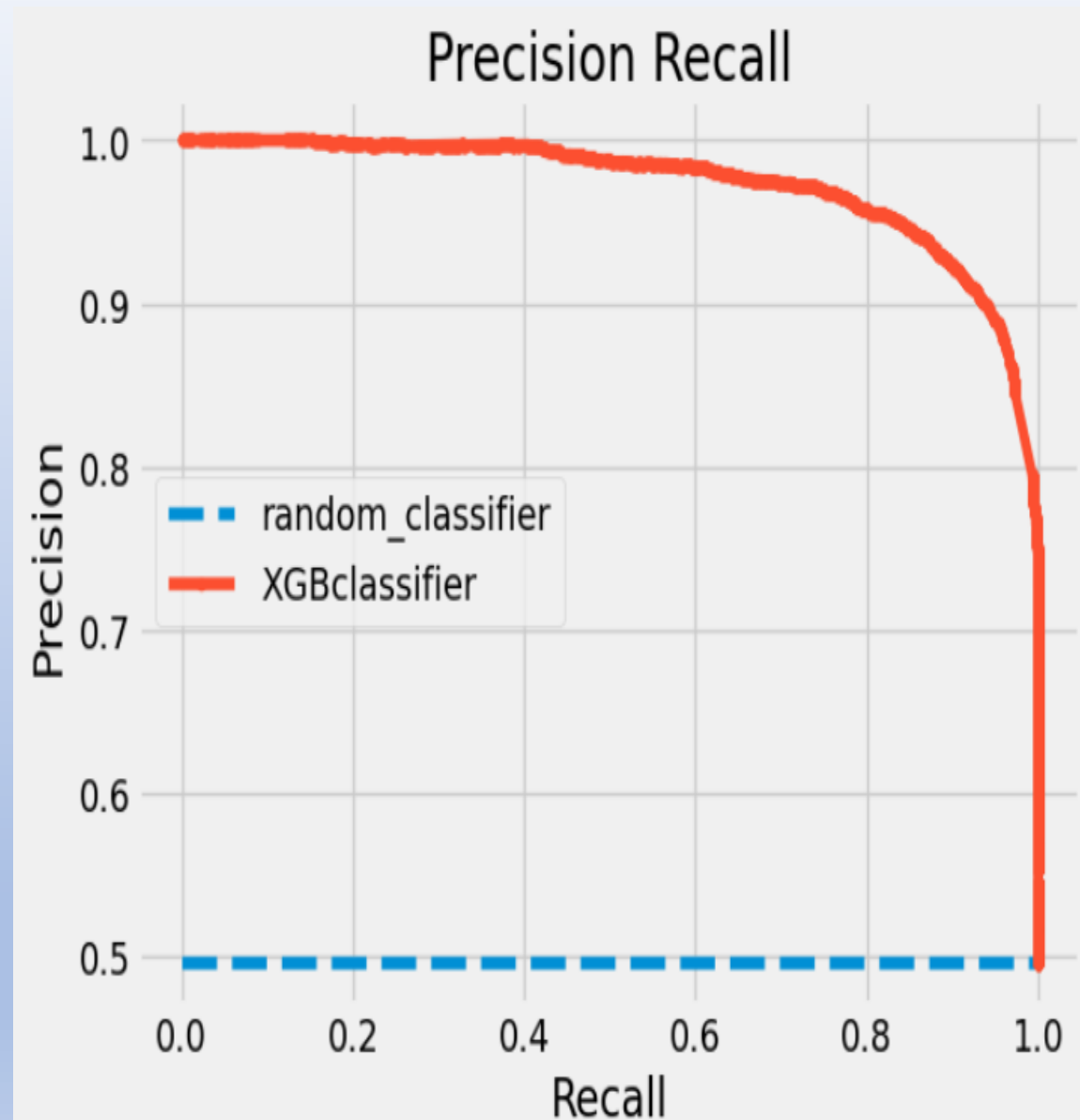
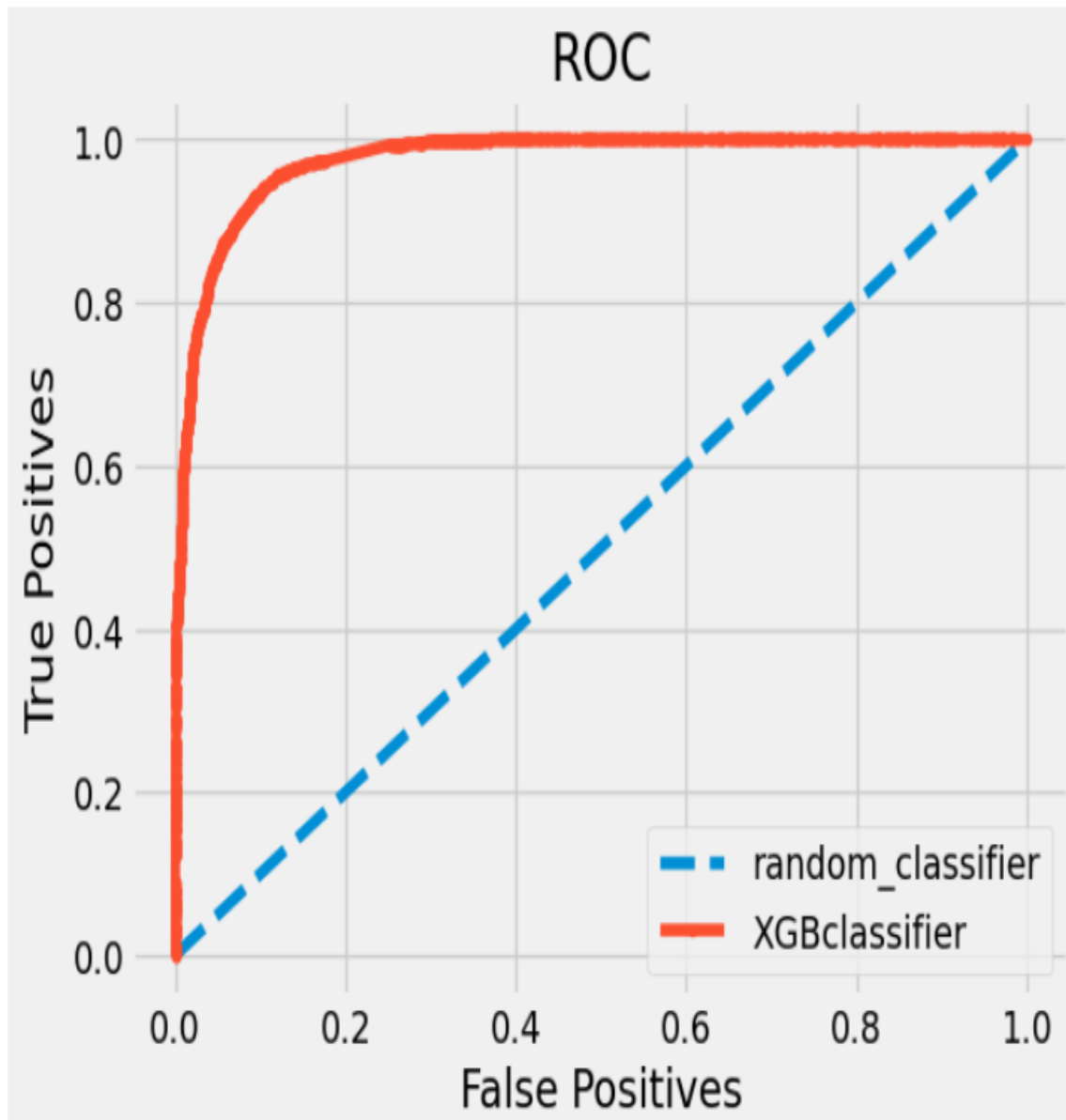
Confusion Matrix:

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.

Model Deployment using Flask

- Deploying a natural language processing (NLP) model on Flask is a way to make the model accessible via a web API.
- Flask is a lightweight web framework for Python that allows for easy creation of web applications and APIs.

The following is a high-level overview of the steps we used to deploy an NLP model on Flask:

1. Train and save the NLP model: Train the NLP model using the desired dataset and save the model to a file. This can be done using a library such as PyTorch or TensorFlow.

2. Create a Flask application: Create a new Flask application and define the routes that the API will handle. These routes will be used to handle incoming requests and return the model's predictions.
3. Load the model: Load the trained NLP model into the Flask application. This can be done using the appropriate library's load function, such as `torch.load()` for PyTorch models.
4. Define the prediction endpoint: Define a route that takes in input data, passes it through the loaded model, and returns the model's predictions. This endpoint will handle the request and response for the API.
5. Start the server: Start the Flask development server to make the API accessible. This can be done using the `app.run()` method.
6. Test the API: Test the API by making requests to the endpoint and checking the returned predictions.

Let's see an example of using the Flask app

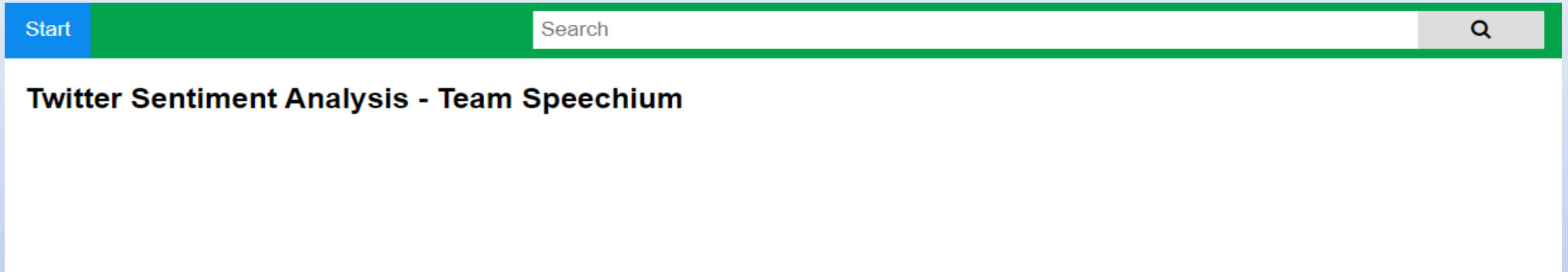
1. Start the server.

```
app.run(debug=False)
```

```
* Serving Flask app '__main__' (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: off

* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

2. Open a web browser and navigate to the application's website using the provided URL: `http://127.0.0.1:5000`.

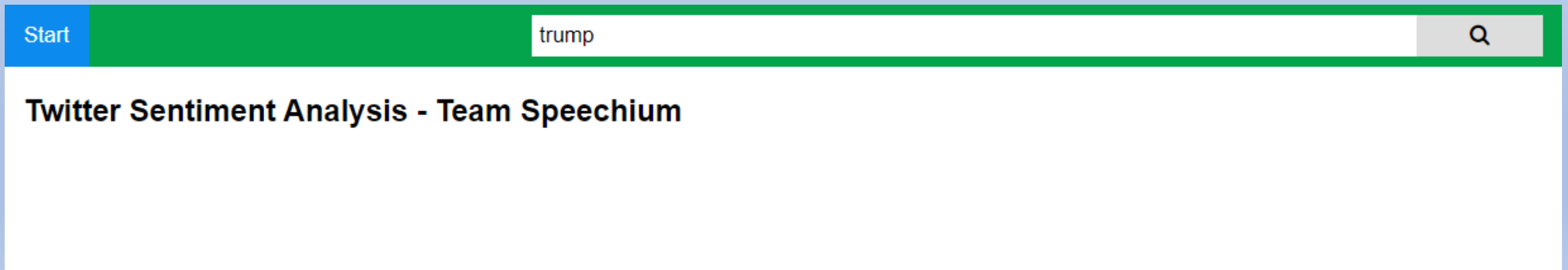


Start

Search

Twitter Sentiment Analysis - Team Speechium

3. Use the search bar to search tweets of your interest.



Start

trump

Twitter Sentiment Analysis - Team Speechium

50 tweets will then be classified as Free Speech or Hate Speech.

```
0      RT @ericareport: Donald Trump being back on Facebook will allow domestic terrorism and white supremacy to spread. Mark Zuckerberg has set a... tweet_text \
1      RT @oneilsteff37: Please leave a red heart for President Trump❤️❤️us https://t.co/DnUTqUpjyP
2      RT @PeteHegseth: What's the first word that comes to mind when you think of President Donald J. Trump?
3      RT @jbrown11871: Remember when Bill Barr said there is no evidence of collusion between Trump and Russia, he lied.
4      RT @TristanSnell: Donald Trump ranting about Fulton County, Georgia DA Fani Willis means that Donald Trump has been told that he's about to...
5      RT @itsJeffTiedrich: holy fucking shit, Donald Trump's tax cuts for the obscenely wealthy are responsible for a QUARTER OF OUR NATION'S ENT...
6      RT @politvidchannel: BREAKING: Calls have been made for Donald Trump to be Tried for 'Mass Murder' over his handling of COVID in 2020
7      RT @TimothyDSnyder: You might remember Manafort's ties to Russia from 2016. He (and Jared Kushner, and Donald Trump, Jr.) met with Russians...
8      RT @TristanSnell: Donald Trump ranting about Fulton County, Georgia DA Fani Willis means that Donald Trump has been told that he's about to...
9      RT @RBReich: Now seems like a good time to remind you that the debt ceiling increased three times under Trump. There was no holding the eco...
10     RT @SenWarren: If Republicans hadn't spent nearly $2 trillion on the Trump tax cuts for giant corporations and hollowed out the IRS to help...
11     RT @RepAdamSchiff: Trump incited an insurrection. And tried to stop the peaceful transfer of power.\n\nHe's shown no remorse. No contrition....
12     RT @itsJeffTiedrich: holy fucking shit, Donald Trump's tax cuts for the obscenely wealthy are responsible for a QUARTER OF OUR NATION'S ENT...
13     RT @bmhochberg: @MarshaBlackburn The debt ceiling needs to be raised to cover Trump's bills, which weren't paid for despite increasing the...
14     RT @Hamletgarcia17: A powerful message from President #Trump As a Cuban American and the grandson of a political prisoner 15-year prison. I...
15     RT @TheTweetOfJohn: The Pennsylvania pizzeria owner who demanded that police at the U.S. Capitol "bring Nancy Pelosi out" to the mob of Don...
16     RT @TimothyDSnyder: In April 2016, I broke the story of Trump and Putin, using Russian open sources. Afterwards, I heard vague intimations...
17     @The_Trump_Train https://t.co/hE02CSIfLr
18     RT @dam7978: @GOP Trying. Did they enter? \nNo. Why? \nBecause as you pointed out, they were apprehended. \nAPPREHENDED. \nA-P-P-R-E-H-E-N-D-E-...
19     RT @robreiner: Pence, Biden, Trump. Only one of these men possessed Highly Classified Top Secret Documents, lied about having them, and ref...
20     RT @RepRaskin: By refusing to seat Adam Schiff, Eric Swalwell and Ilhan Omar on their committees while seating compulsive liar and conman G...
21     @marklevinshow I'm kinda thinking that trump.snd Obama declassified before the letf office. Biden didn't have that oppertunity!
22     RT @MayoIsSpicyy: Donald Trump is the dumbest and most corrupt president that America has ever had.
23     RT @RBReich: Now seems like a good time to remind you that the debt ceiling increased three times under Trump. There was no holding the eco...
24     RT @DavidGr07837209: Donald Trump's 2024 "campaign" is falling apart already https://t.co/rEt6Yu2GYI via @PalmerReport
25     RT @Pismo_B: President Donald Trump allowed back on Facebook and Instagram, Meta announces https://t.co/QEBW1NkinR
26     RT @TimothyDSnyder: In April 2016, I broke the story of Trump and Putin, using Russian open sources. Afterwards, I heard vague intimations...
27     @sdnativejoe @RitchieTorres Are you naturally that stupid or do you practice? The blatant lies Schiff told about Tr... https://t.co/kZmwTsDEnP
28     @dumdum989 @The_Trump_Train Millions of people took the vaccine. Enjoy your life, stop reading the twitter crap.... https://t.co/o55ESH8GjM
29     @mar80964942 @MarshaBlackburn Actually it was with Trump.\nhttps://t.co/FFMiT9ADU1
30     RT @tedlieu: President Biden and VP Pence did not intend to take classified documents and then refuse to give them back. But former Preside...
31     @TomFitton I honestly don't even want Trump anymore.
32     RT @krassenstein: Share This Post if you know:\n\nCOVID vaccines WORK 👍\nClimate change is REAL 👍\nDr. Fauci is a HERO 👍\nDonald Trump is a CROO...
33     RT @TimothyDSnyder: Manafort had to resign as Trump's campaign manager in August 2016 when news broke that he had received $12.7 million in...
34     RT @Pismo_B: @POTUS President Trump signed the executive order in 2020 that capped the price of insulin at $35 for Medicare enrollees, Bide...
35     @lancewallnau Zuckerberg is a horrible, unethical person. I hope Mr. Trump stays true to his word and is exclusively on TRUTH SOCIAL!
36     @ericswalwell @thereidout Oh my god will you please go cry somewhere else shut up and move on already!you lied and... https://t.co/7oo0JF3CBv
37     RT @GregBrinsonmtb: @RaheemKassam @Rich_Cranium4 Trump do you still believe your Warp Speed Clot Shot saved 100 million he better come clean
38     It's unclear whether Donald Trump will again become active on Facebook and Instagram. Truth Social is currently the... https://t.co/TsKHwjQW7r
39     RT @thejackhopkins: Kevin McCarthy's acts, carried out on behalf of Putin, via Donald Trump and the Putin supporting republican party, will...
40     RT @robreiner: Pence, Biden, Trump. Only one of these men possessed Highly Classified Top Secret Documents, lied about having them, and ref...
41     @dspov @RBReich The republican party never show's concern until a democrat becomes president, then miraculously the... https://t.co/PzQIrL3loi
42     @The_Trump_Train YES, because it WAS a MURDER
43     RT @politvidchannel: BREAKING: Calls have been made for Donald Trump to be Tried for 'Mass Murder' over his handling of COVID in 2020
44     @The_Trump_Train Yes
45     @RonFilipkowski Trump Grump will be in jail by then, hopefully. 🤔👉
46     RT @PlusGiftcards: @JagodkaJason @Goodbye_Jesus Antichrist Donald Trump is about to get the fake world peace treaty confirmed he's been pla...
47     Would have never happened under Trump. https://t.co/pt9rHZFzg1
48     @mrbackhand @StatesPoll https://t.co/j1QvBaaAvJ
49     @poneilinOttawa @ChanGardiner @TimothyDSnyder @TuffTiffResists @nytimes The @nytimes that helped Trump win? The... https://t.co/Z829o6LDWX
```

The model predicted:

- 48 Free Speech tweets and
- 2 Hate Speech tweets

```
0    48  
1     2  
Name: prediction, dtype: int64
```

		created_at	tweet_id
0	2023-01-26	01:28:27+00:00	1618420569413914626
1	2023-01-26	01:28:27+00:00	1618420568935772160
2	2023-01-26	01:28:27+00:00	1618420568805769216
3	2023-01-26	01:28:26+00:00	1618420568021422080
4	2023-01-26	01:28:26+00:00	1618420567757160449
5	2023-01-26	01:28:26+00:00	1618420566951890944
6	2023-01-26	01:28:26+00:00	1618420566473736193
7	2023-01-26	01:28:26+00:00	1618420566226268160
8	2023-01-26	01:28:26+00:00	1618420566201085954
9	2023-01-26	01:28:26+00:00	1618420565932670976
10	2023-01-26	01:28:26+00:00	1618420565144121345
11	2023-01-26	01:28:26+00:00	1618420564884094976
12	2023-01-26	01:28:25+00:00	1618420564246532096
13	2023-01-26	01:28:25+00:00	1618420564049428480
14	2023-01-26	01:28:25+00:00	1618420562908573698
15	2023-01-26	01:28:25+00:00	1618420562589798403
16	2023-01-26	01:28:25+00:00	1618420562585587712
17	2023-01-26	01:28:25+00:00	1618420562493337600
18	2023-01-26	01:28:25+00:00	1618420562413654018
19	2023-01-26	01:28:25+00:00	1618420561289543683
20	2023-01-26	01:28:25+00:00	1618420561209856001
21	2023-01-26	01:28:25+00:00	1618420560656232449
22	2023-01-26	01:28:24+00:00	1618420556982018048
23	2023-01-26	01:28:24+00:00	1618420556948451328
24	2023-01-26	01:28:24+00:00	1618420556612915202
25	2023-01-26	01:28:24+00:00	1618420556612915200
26	2023-01-26	01:28:24+00:00	1618420556109590529
27	2023-01-26	01:28:23+00:00	1618420555585314816
28	2023-01-26	01:28:23+00:00	1618420555237167104
29	2023-01-26	01:28:23+00:00	1618420554725462016
30	2023-01-26	01:28:23+00:00	1618420554536718337
31	2023-01-26	01:28:23+00:00	1618420554436050948
32	2023-01-26	01:28:23+00:00	1618420554347970562
33	2023-01-26	01:28:23+00:00	1618420554176032770
34	2023-01-26	01:28:23+00:00	1618420553978904580
35	2023-01-26	01:28:23+00:00	1618420553874018307
36	2023-01-26	01:28:23+00:00	1618420553651720192
37	2023-01-26	01:28:23+00:00	1618420553244905473
38	2023-01-26	01:28:23+00:00	1618420552783335424
39	2023-01-26	01:28:23+00:00	1618420552770936844
40	2023-01-26	01:28:23+00:00	1618420552141795329
41	2023-01-26	01:28:22+00:00	1618420551231623169
42	2023-01-26	01:28:22+00:00	1618420550992564225
43	2023-01-26	01:28:22+00:00	1618420550824787969
44	2023-01-26	01:28:22+00:00	1618420550334054400
45	2023-01-26	01:28:22+00:00	1618420549503549440
46	2023-01-26	01:28:22+00:00	1618420549394518018
47	2023-01-26	01:28:22+00:00	1618420548308205568
48	2023-01-26	01:28:22+00:00	1618420547876184067
49	2023-01-26	01:28:21+00:00	1618420547251220480

We can also identify which are the Hate Speech tweets

	prediction
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	1
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	0
29	0
30	0
31	0
32	0
33	0
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	1
46	0
47	0
48	0
49	0

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.

Thank You