NLP-Powered Surveillance Against Online Hate

Ashwini Narayana Shetty
Rohan Ranjit Kamath
Venkatesh Murugesh
Shivani Anil Neharkar
Tejaswini Shinde

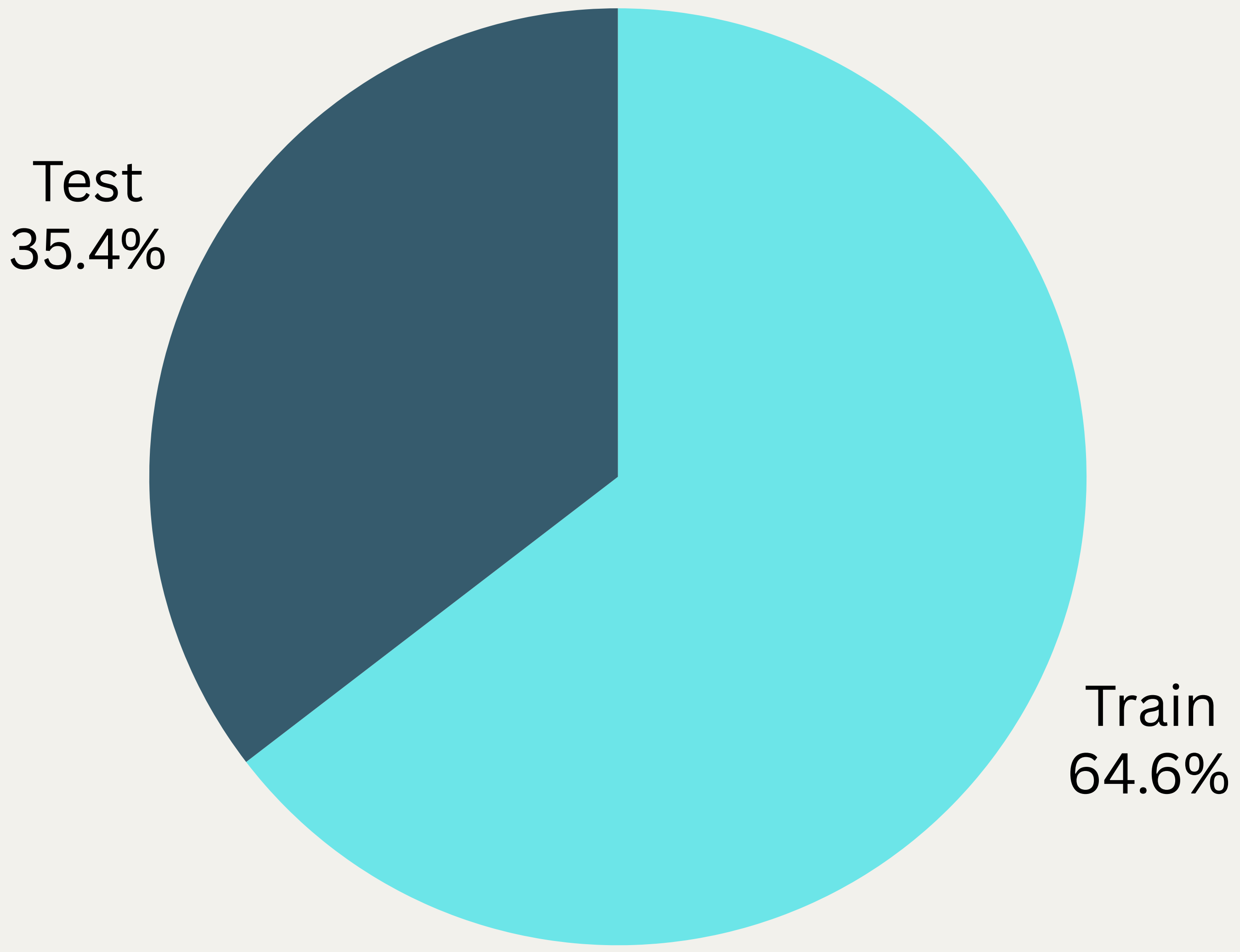
INTRODUCTION

- **The Problem:** Hate speech is widespread on social media, appearing in both explicit and implicit forms, leading to discrimination and violence.
- **Our Goal:** Develop a machine learning model to detect and classify hate speech, offensive language, and neutral content with consistent accuracy across various social media platforms.
- **Our Approach:** To employ advanced NLP techniques to analyse social media content, implementing a multi-model strategy with Deep Neural Networks with BERT and TF-IDF embeddings to improve accuracy.

ABOUT THE DATASET

train.csv - contains 31000+ tweets
test.csv - contains 17000+ tweets

Fig 1: Training Test Distribution



Attributes of the dataset

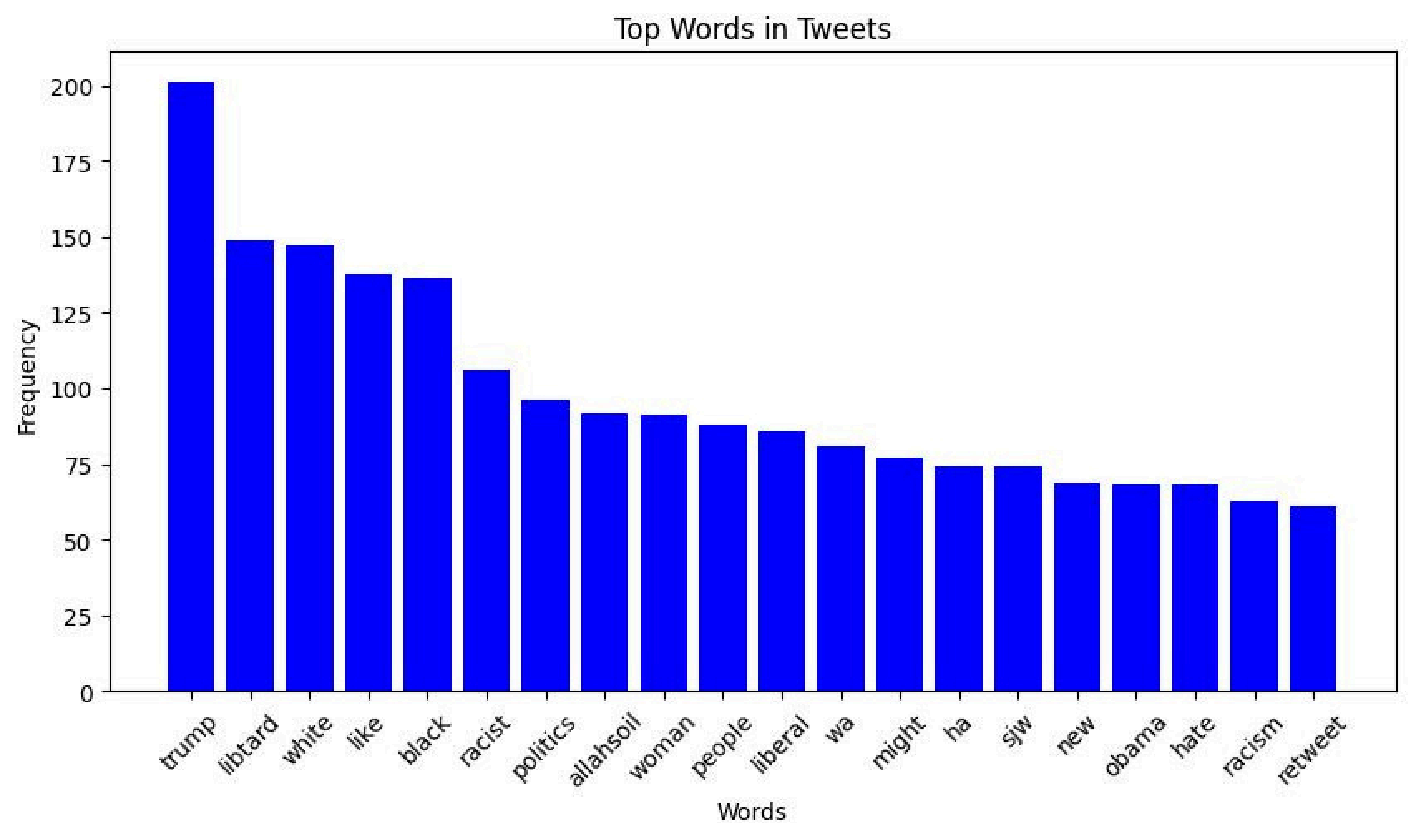
Text

The text of the tweet

Target

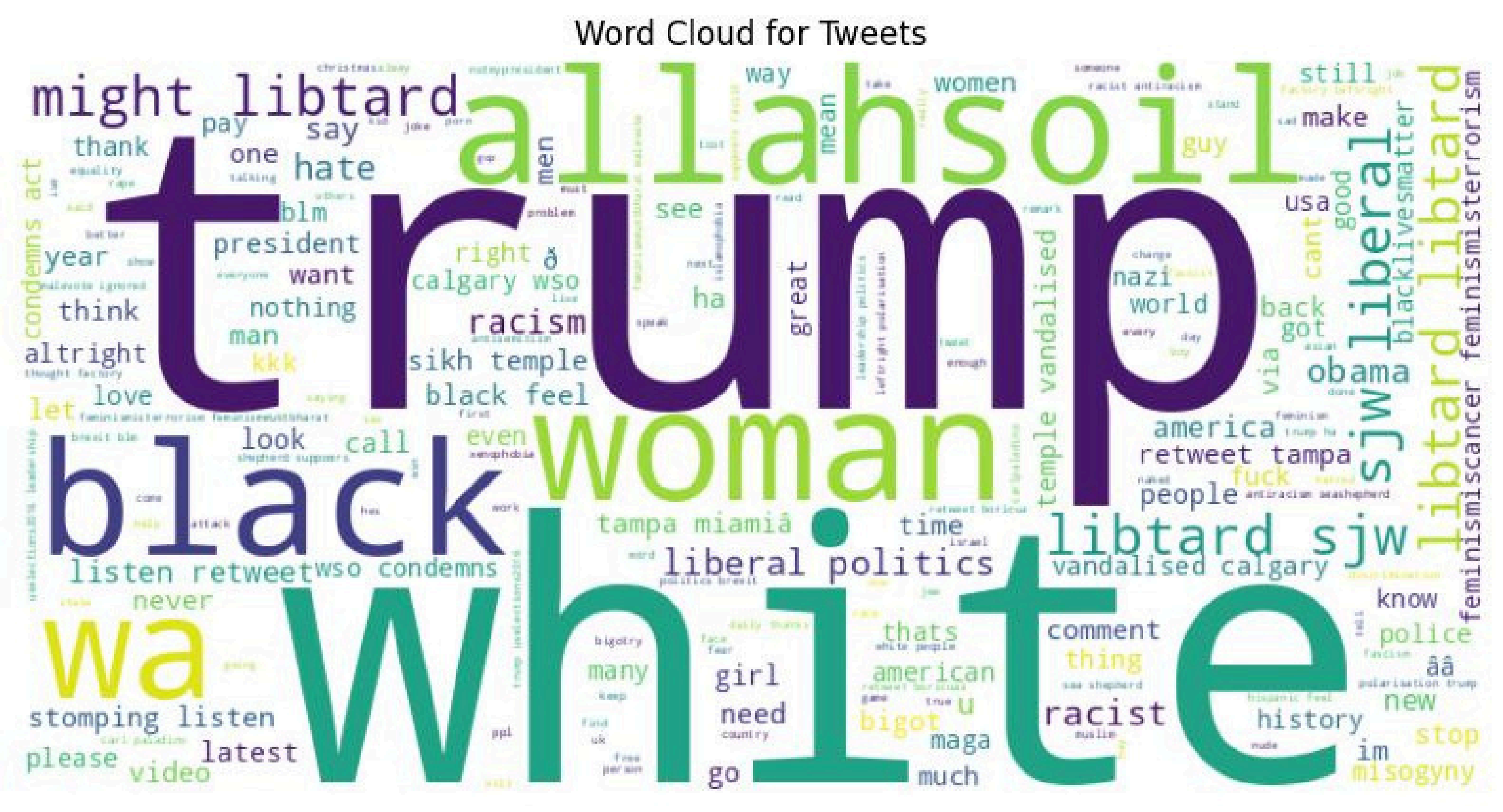
denotes whether a tweet considered “hateful” or not (1/0)

ABOUT THE DATASET



Top 20 Words

ABOUT THE DATASET

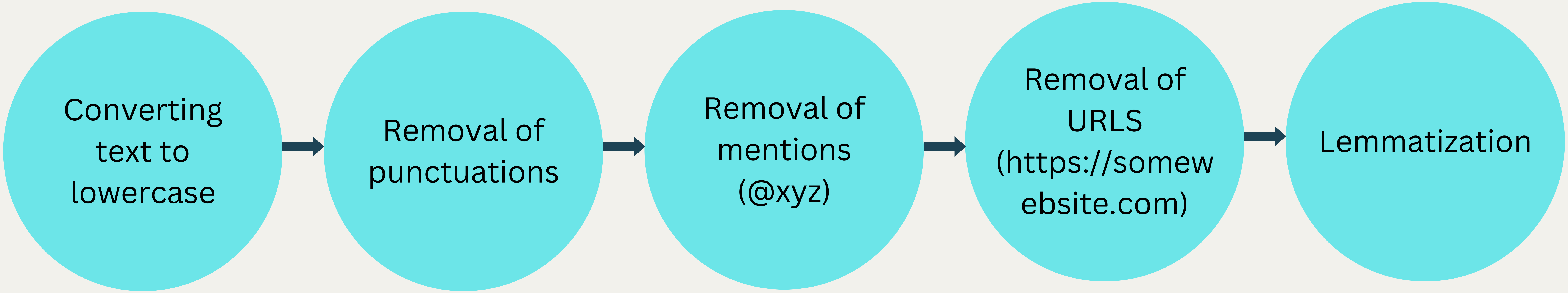


Word Cloud for Tweets

Word Cloud

METHODOLOGY

A) DATA PREPROCESSING



Converting text to lowercase

Removal of punctuations

Removal of mentions (@xyz)

Removal of URLS (https://somewebsite.com)

Lemmatization

B) TEXT TO VECTOR CONVERSION

- Distilled BERT
- TF-IDF vectorization

C) CLASSIFICATION MODELS

- Traditional ML Algorithms (more on the next slides)

EMBEDDINGS USED

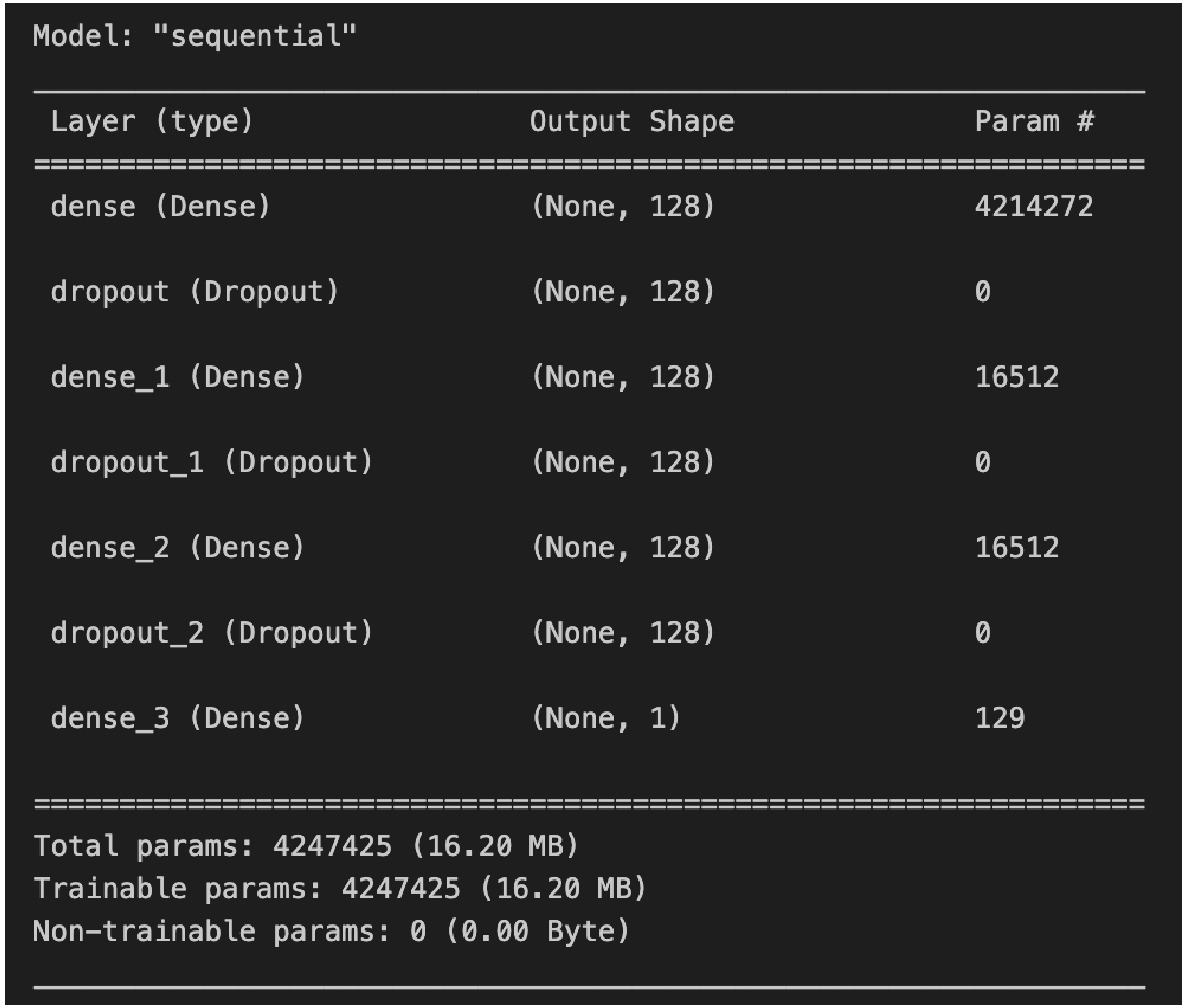
Sr. No.	Embeddings
01	Distilled BERT
02	TFIDF
03	BERT + TFIDF

ML MODELS USED

Sr. No.	ML Model
01	Deep Neural Networks
02	Logistic Regression
03	Random Forest

*We are going to mix and match the embeddings and ML models

MODEL1: DEEP NEURAL NETWORKS



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               4214272

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 128)               16512

 dropout_1 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 128)               16512

 dropout_2 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 1)                 129

=================================================================
Total params: 4247425 (16.20 MB)
Trainable params: 4247425 (16.20 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

NN Layers:

- 1 input layer
- 2 hidden layers
- 1 output layer (1 node, sigmoid function for classification)

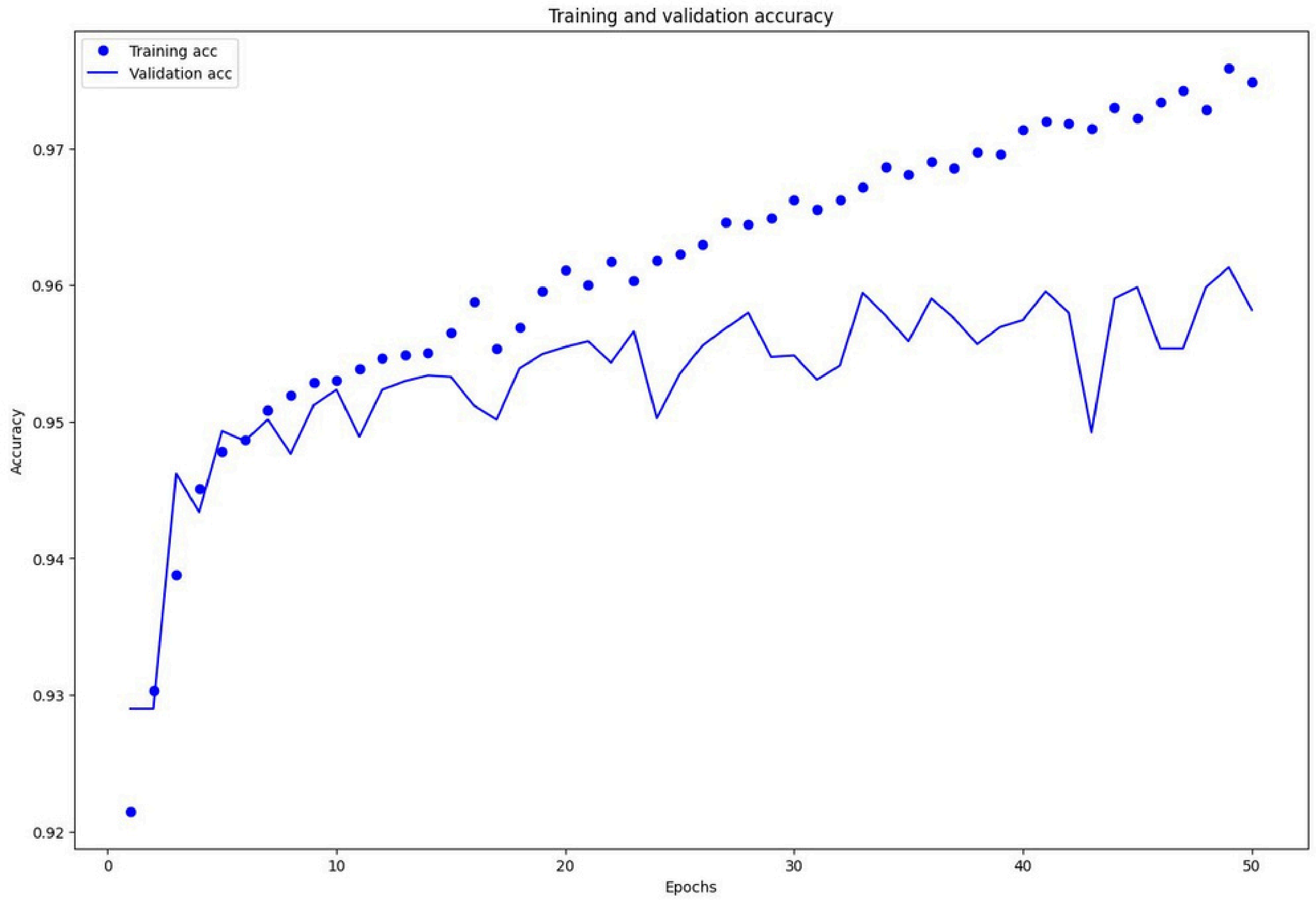
MODEL 2: LOGISTIC REGRESSION

MODEL 3: RANDOM FOREST

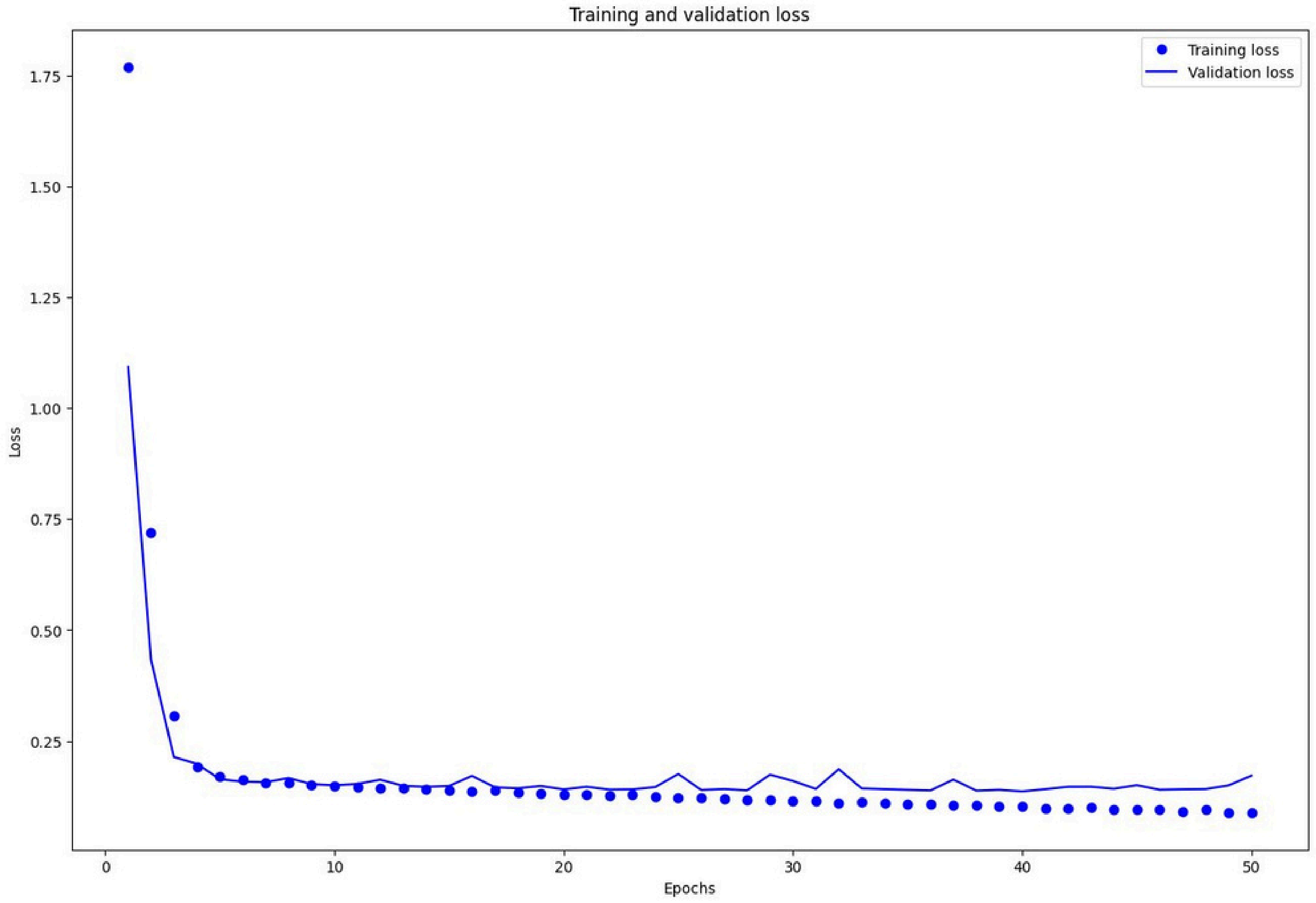
*Results in the upcoming slides.

DNN OBSERVATIONS

Improvement of Accuracy on Training Data



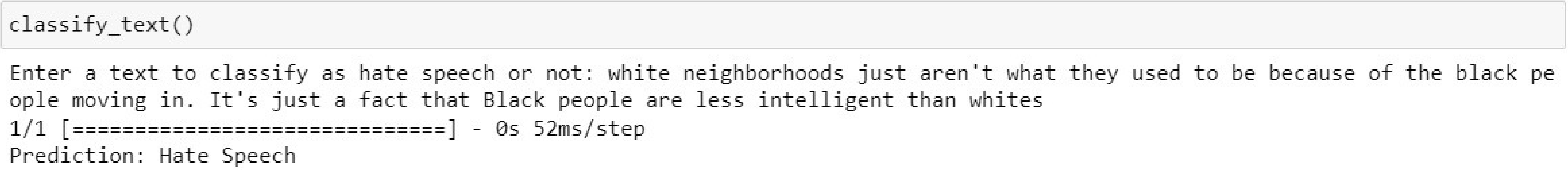
Training and Validation Loss



RESULTS

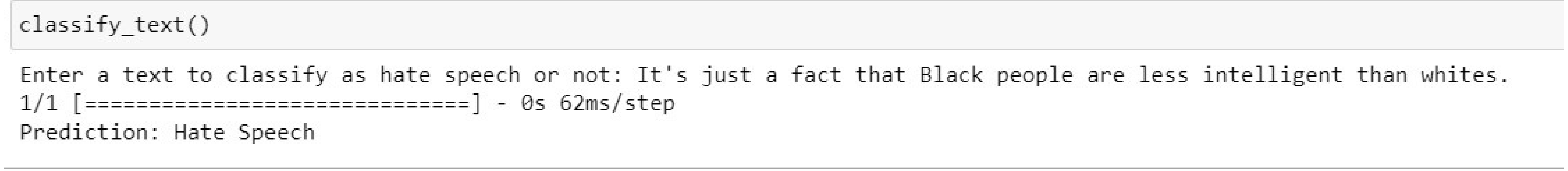
Embeddings	Model	F1-Score
Distilled BERT	DNN	65%
	LG	61%
	RF	40%
TF-IDF	DNN	65%
	LG	42%
	RF	64%
BERT + TF-IDF	DNN	**74%**
	LG	62%
	RF	39%

MODEL OUTPUT



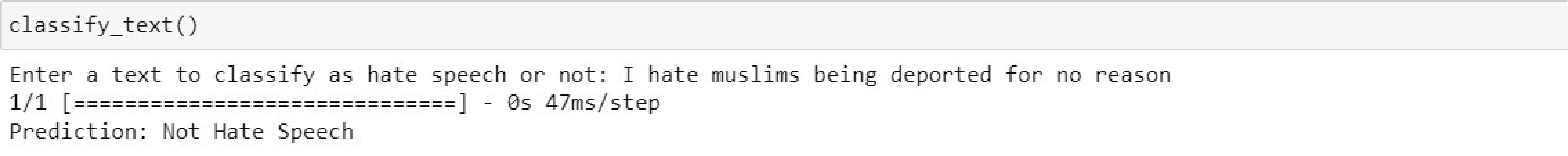
```
classify_text()
```

```
Enter a text to classify as hate speech or not: white neighborhoods just aren't what they used to be because of the black pe
ople moving in. It's just a fact that Black people are less intelligent than whites
1/1 [==============================] - 0s 52ms/step
Prediction: Hate Speech
```



```
classify_text()
```

```
Enter a text to classify as hate speech or not: It's just a fact that Black people are less intelligent than whites.
1/1 [==============================] - 0s 62ms/step
Prediction: Hate Speech
```



```
classify_text()
```

```
Enter a text to classify as hate speech or not: I hate muslims being deported for no reason
1/1 [==============================] - 0s 47ms/step
Prediction: Not Hate Speech
```

DEMO

THANK YOU!