

# Twitter Sentiment Analysis

Bhat Sachin | U2123512F

Nalin Sharma | U2121904E

DSF1 Group 11



# Practical Motivation

- A common occurrence on the internet
- Censorship issues
- Strong connection between hate speech and actual hate crime (HateLab project - Cardiff University)
- Early identification --> Enabling outreach programmes
- NLP Research on hate speech has been limited



# Christchurch mosque shootings

A contemporary example of hate speech materialising into hate crime.

- A mass shooting that occurred in Christchurch on 15 March 2019, leaving 50 people dead and dozens others wounded.
- Shooter posted about his plans on 8chan: "time to stop shitposting and time to make a real life effort".



# Problem Formulation



*Problem to be addressed:  
Effective implementation of Data Science and  
Natural Language Processing (NLP) concepts to  
find the best model to detect hate speech in  
tweets.*

*Guiding Question: How can we effectively  
detect hate speech in tweets?*

# The Dataset

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disappointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urđ□□±!!! đ□□□đ□□□đ□□□đ□□□ đ□□!đ□□!đ□□!
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ□!
7	8	0	the next school year is the year for exams.đ□□ can't think about that đ□□ #school #exams #hate #imagine #actorslife #revolutionschool #girl
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â□!
9	10	0	@user @user welcome here ! i'm it's so #gr8 !
10	11	0	â□□ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog #silver #gold #forex
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking #values #love #
12	13	0	i get to see my daddy today!! #80days #gettingfed
15	16	0	ouch...junior is angryđ□□□#got7 #junior #yugyoem #omg
16	17	0	i am thankful for having a paner. #thankful #positive

*Snippet from Train set*

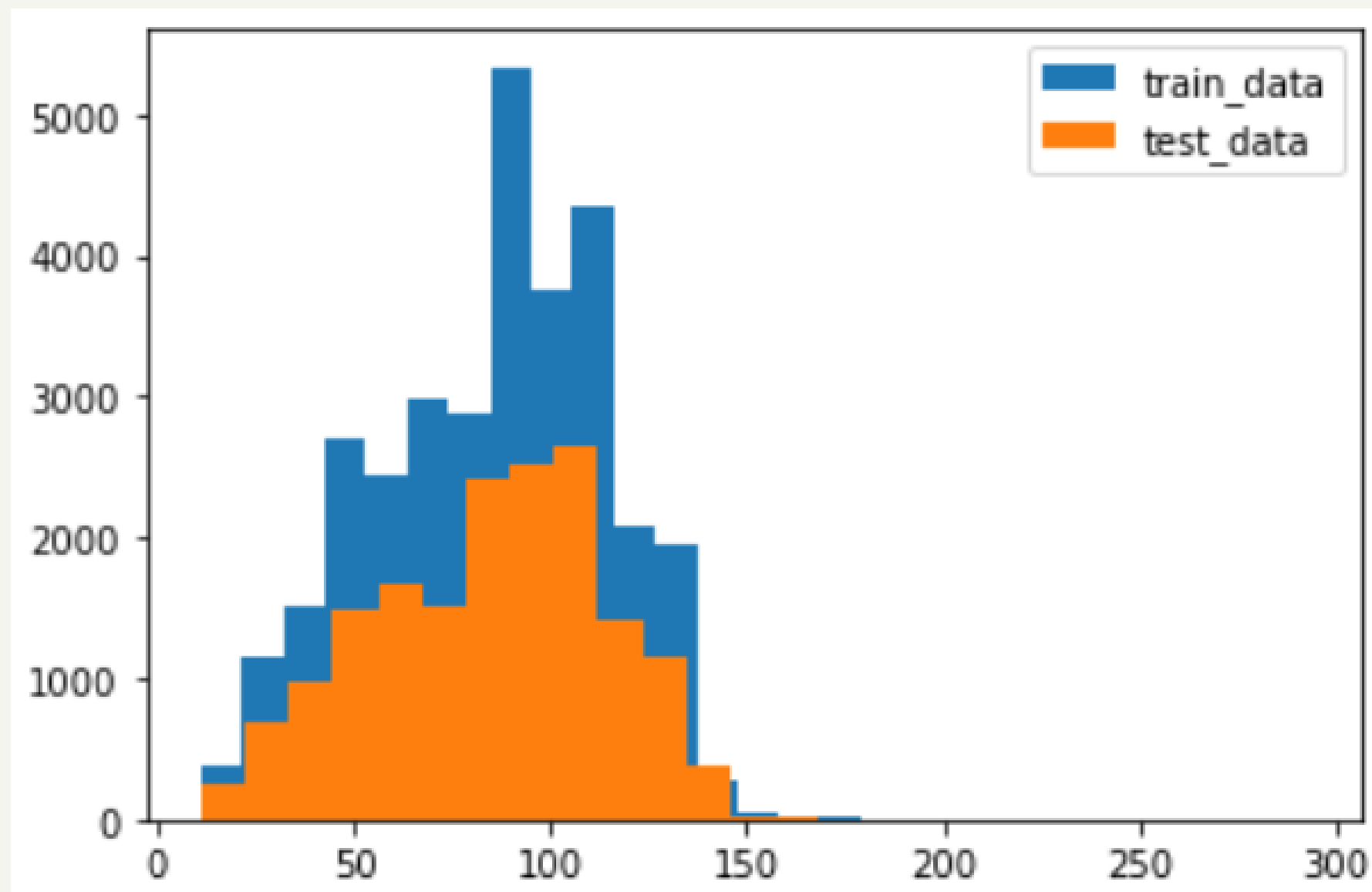
```
print("Data type : ", type(train))
print("Data dims (train): ", train.shape)
print("Data dims (test): ", test.shape)
```

```
Data type : <class 'pandas.core.frame.DataFrame'>
Data dims (train): (31962, 3)
Data dims (test): (17197, 2)
```

**Train set:** 31,962 tweets in 3 columns

**Test set:** 17,197 tweets in 2 columns

# Exploratory Data Analysis



- Similarity between distribution of the length of tweets of test and train data
- Overall shape is similar => well-distributed train-test split.

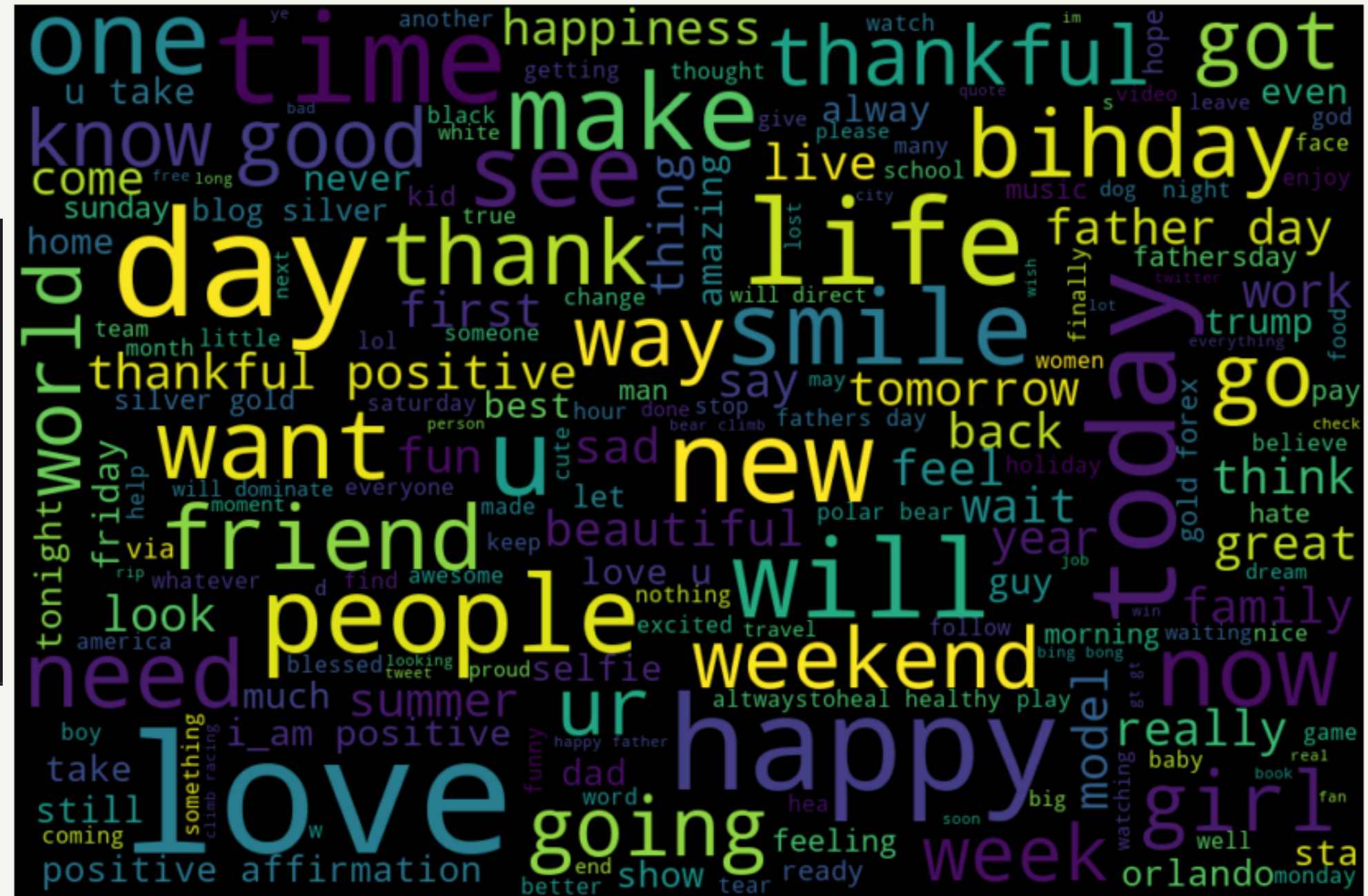


# Wordcloud Representation

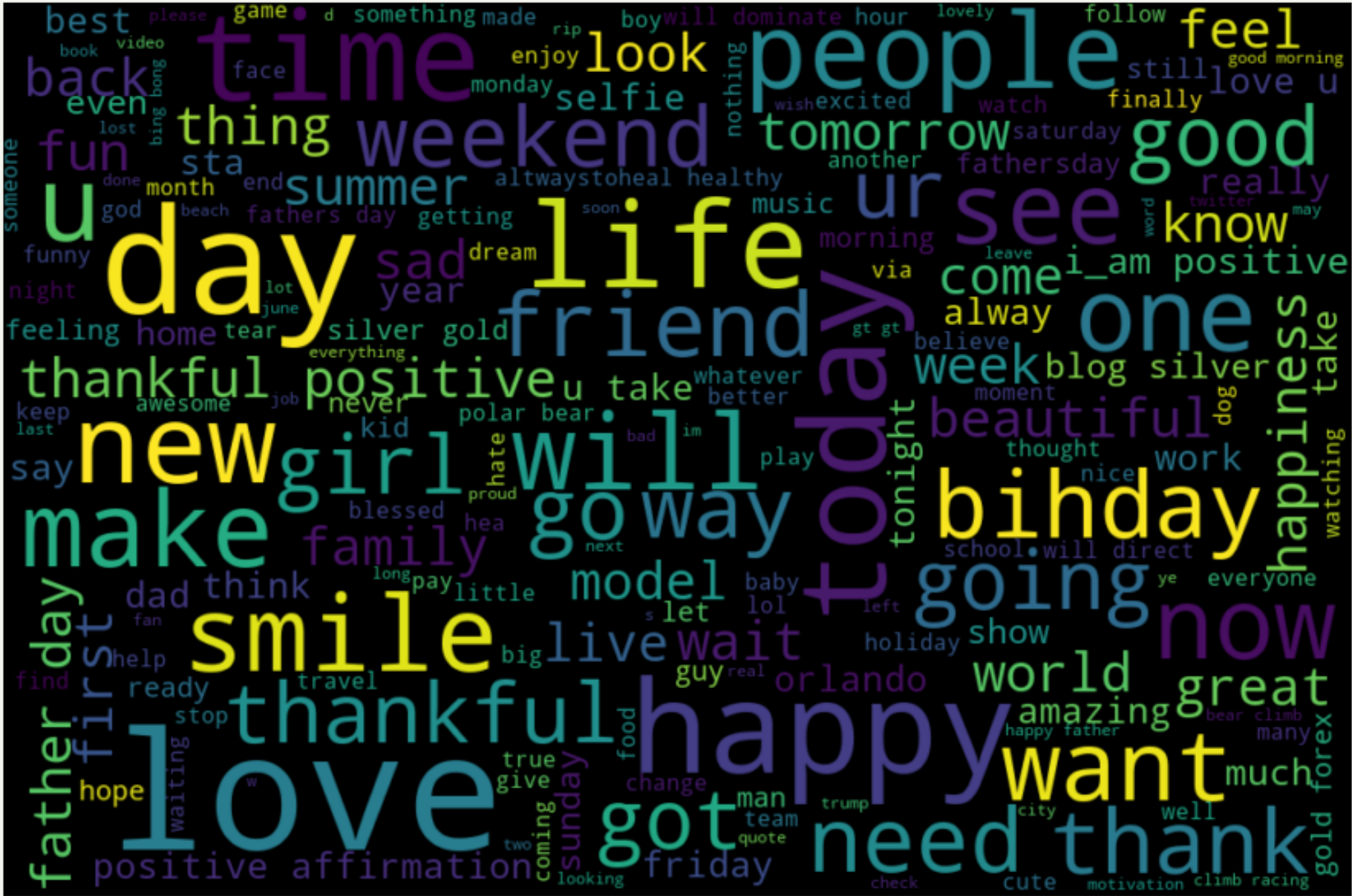
```
words = ' '.join([token for token in data_clean['Cleaned Tweet']])

from wordcloud import WordCloud
wordcloud = WordCloud(width= 980,
                      height= 650,
                      random_state=21,
                      max_font_size=120).generate(words)

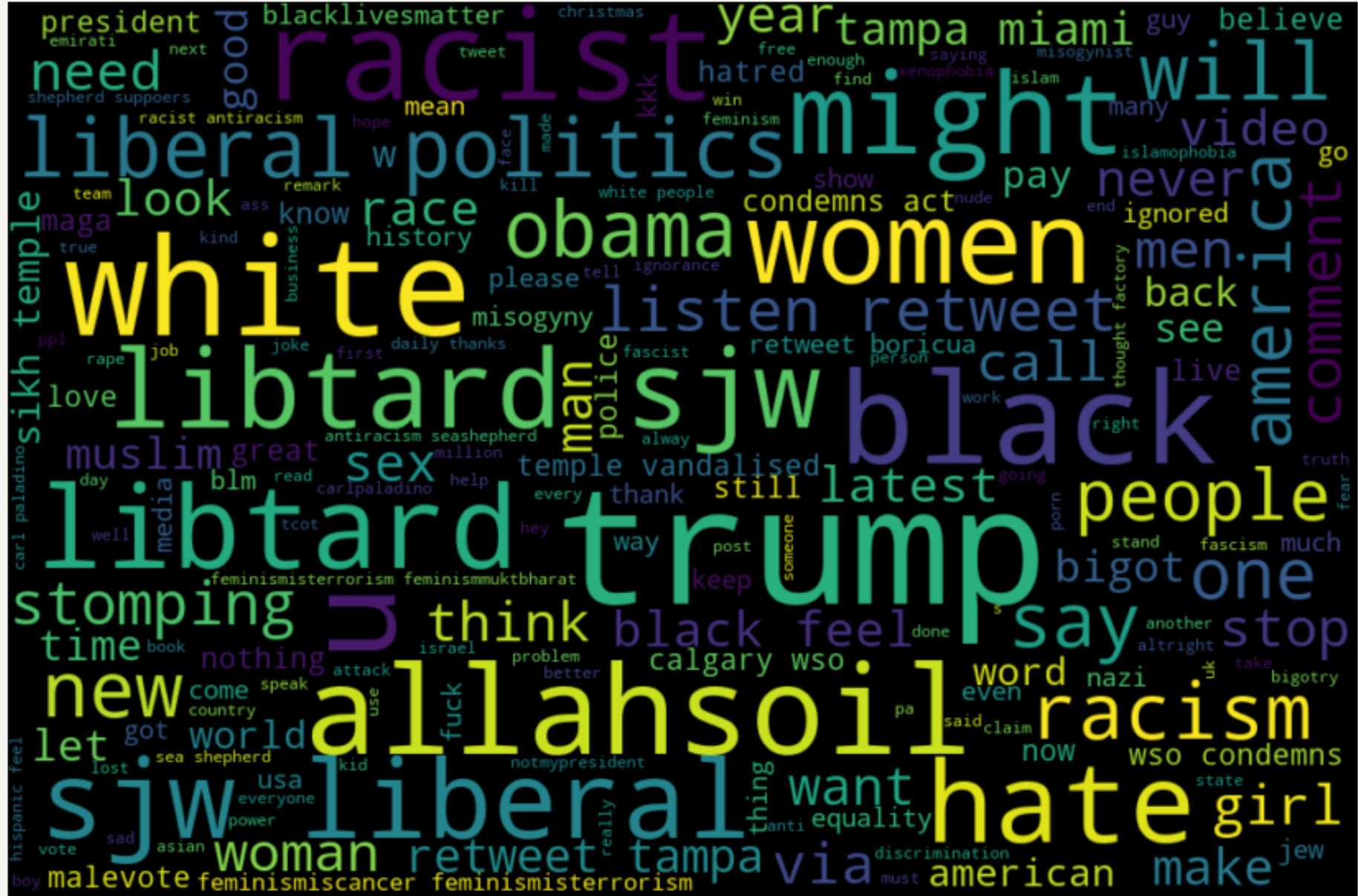
plt.figure(figsize=(15, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



# Positive



# Negative





# Data Cleaning

1

Renaming & dropping  
redundant columns

2

Reindexing

3

Removing non-ascii  
values

4

@user and other  
anomalies

# data\_clean

	label	Cleaned Tweet
1	0	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disappointed #getthanked
3	0	bihday your majesty
4	0	#model i love u take with u all the time in ur!!!
5	0	factsguide: society now #motivation
...	...	...
31958	0	ate isz that youuu?
31959	0	to see nina turner on the airwaves trying to wrap herself in the mantle of a genuine hero like shirley chisolm. #shame #imwithher
31960	0	listening to sad songs on a monday morning otw to work is sad
31961	1	#sikh #temple vandalised in in #calgary, #wso condemns act
31962	0	thank you for you follow
31962 rows × 2 columns		

# Text Normalisation

1

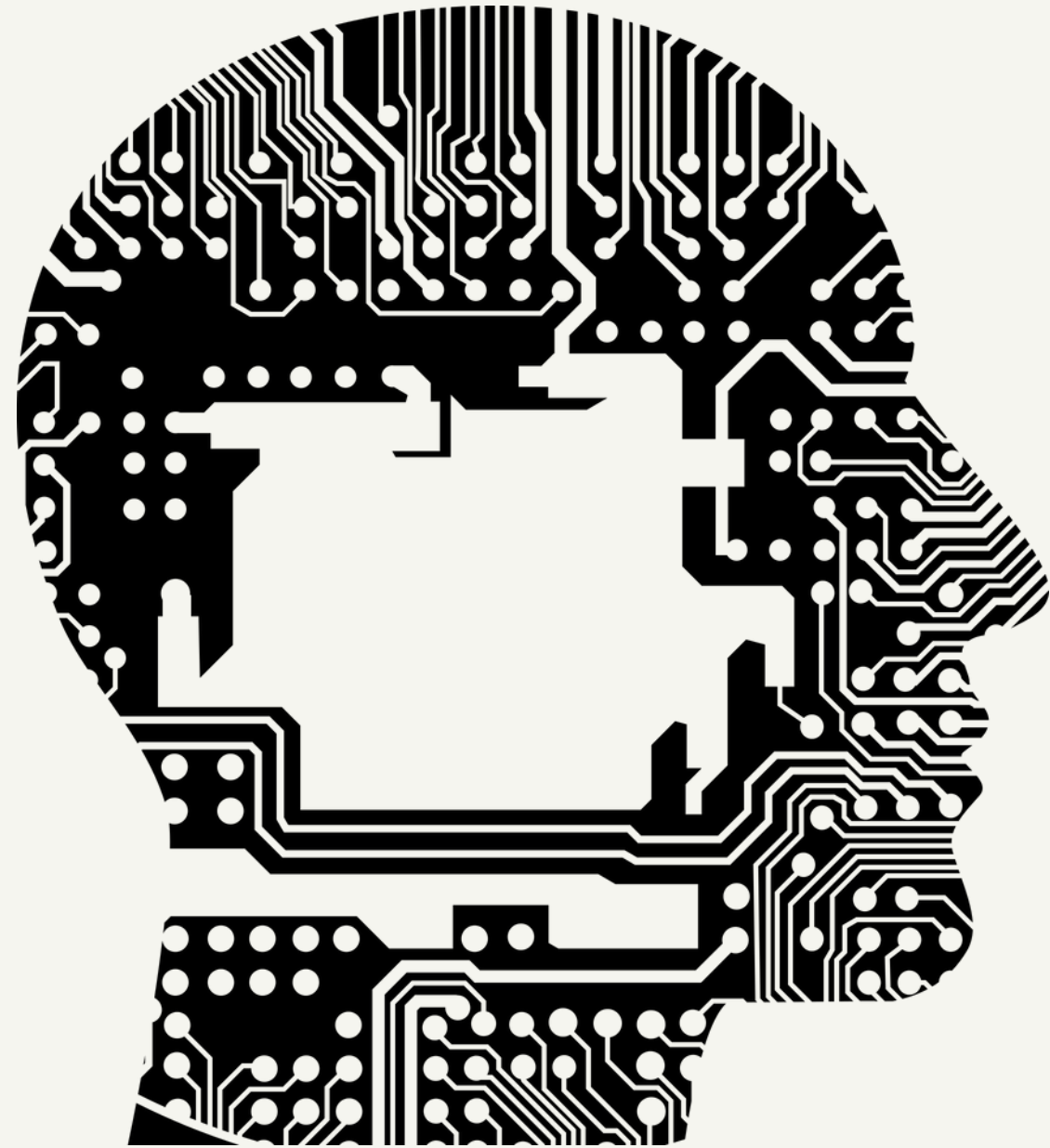
Tokenise tweets

```
[14]: tweet_tokens = data_clean['Cleaned Tweet'].apply(lambda x: x.split()) # tokenising tweets
```

2

Normalise tweets

```
In [16]: from nltk.stem.porter import *  
         pStemmer = PorterStemmer()  
         tweet_tokens = tweet_tokens.apply(lambda x: [pStemmer.stem(i) for i in x])
```



# Feature Extraction



# Key Features

1

Bag-Of-Words

2

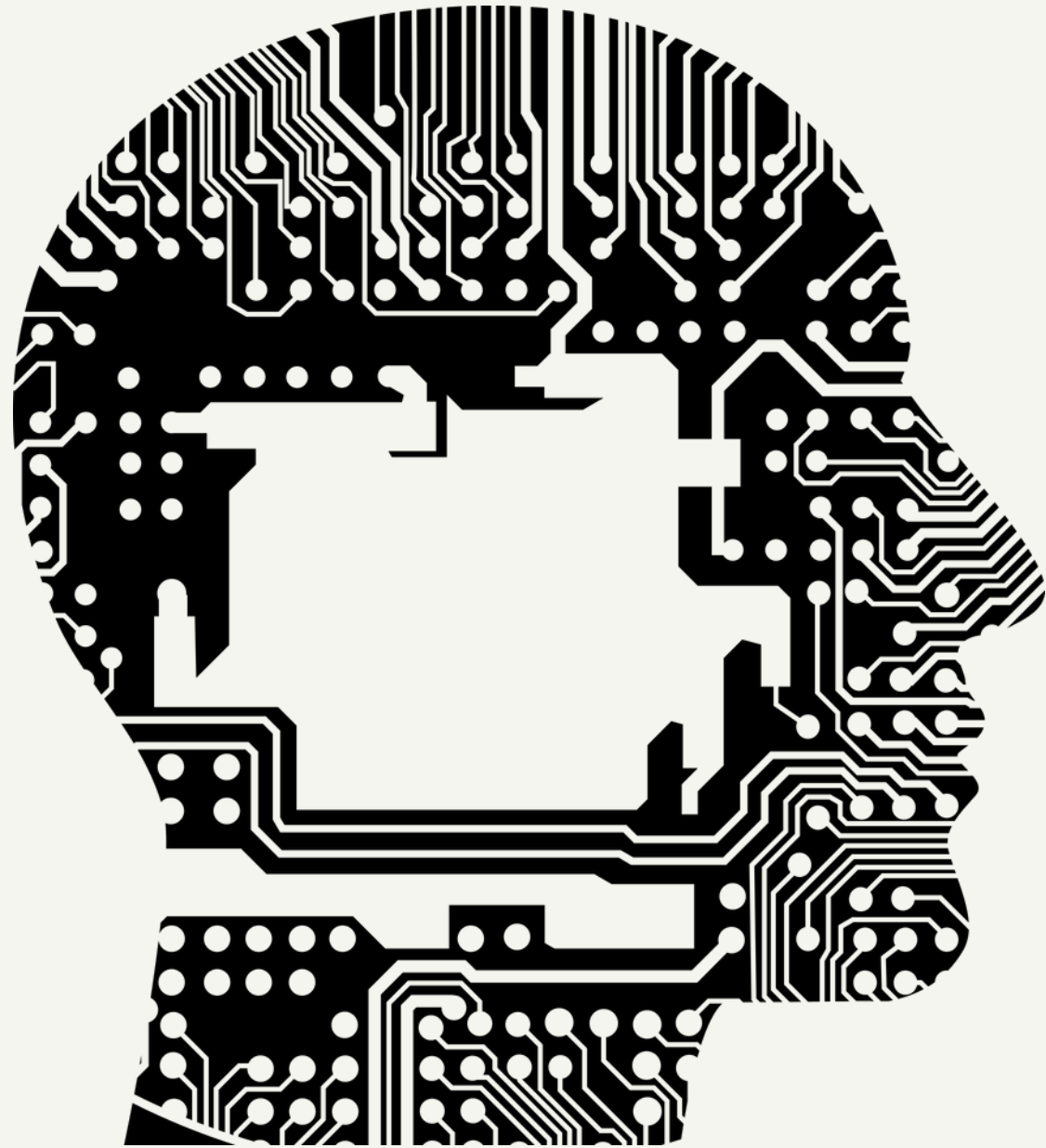
TF-IDF

3

Word2Vec

4

Doc2Vec



# Machine Learning

Models used:

1. Support Vector Machine
2. Logistic Regression
3. Random Forest
4. XGBoost

# Performance Analysis

The average F1 Score for SVM is 0.418

The average F1 Score for Logs Regression is 0.441

The average F1 Score for RF is 0.371

The average F1 Score for XGBoost is 0.493



0.493

Average F1 Score for XGBoost

0.621

Highest F1 Score achieved by  
XGBoost via Word2Vec

~0.750

Expected F1 Score after Tuning



# General Structure for Hyperparameter Tuning

01

Chose a relatively high learning rate

Usually LR = 0.3 at this stage

02

Tune and update tree-specific parameters

E.g. max\_depth,  
min\_child\_weight,  
subsample,  
colsample\_bytree

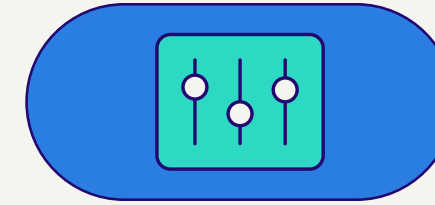
03

Tune and update the learning rate.

04

Tune and update 'Gamma' to prevent overfitting.

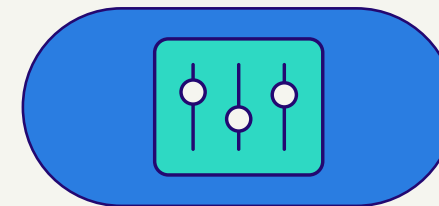
# Conclusion



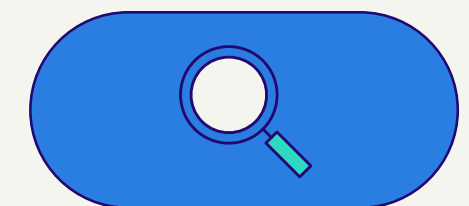
XGBoost as the  
best model



Word2Vec as the  
best feature

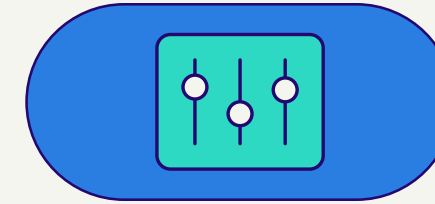


Hyperparameter  
Tuning for  
optimisation



Useful to analyse  
hate crime  
motives

# Recommendations / Possible Extensions



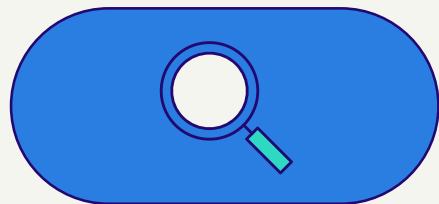
## Hyperparameter Tuning

Optimising model performance



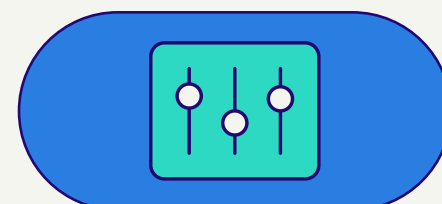
## Trying out other binary classifiers

E.g. simple decision tree, naive bayes etc.



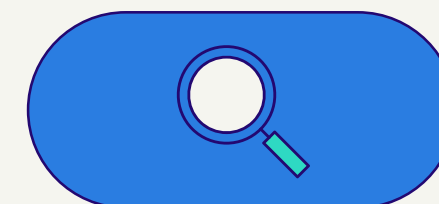
## Examining specific sentiments

Such as depression, joy, anger.



## Optimising Precision and Recall

To improve time and space complexity



## Visualisation

Use of GraphViz to illustrate results



## Social media posts

Implementing the same techniques on platforms such as Instagram to gain a deeper insight



Thank you!