PROBLEM STATEMENT

To classify offensive tweets from non-offensive ones from a repository of 32k tweets.

Relevance

- There is a substantial amount of hateful content on the web that incites violence and aggressive behavior.
- To curb violence and criminal activity in the society, it is important to mitigate hateful/racist information.
- Twitter can be a proxy measurement of racism, and research is being conducted to better understand how racist tweets impacts public health[1].

Data introduction

- A collection of 32K tweets labeled as hatred (racist/sexist) or non hatred which is part of training data[2].
- The aim of the project is to train and compare both classical models (logistic/random trees/gradient boost) as well as neural network models. The most accurate model is used to predict offensive tweets from test data.

[1]https://news.furman.edu/2021/01/22/words-matter-study-looks-at-tweets-impact-on-health/.
[2]https://www.kaggle.com/arkhoshghalb/detecting-hate-tweets/data.

Who will benefit?

- Various governments, law enforcement agencies and non-profit NGOs fighting against racism will be interested in accounts which display offensive tweets.

What factors likely contribute to customer subscription?

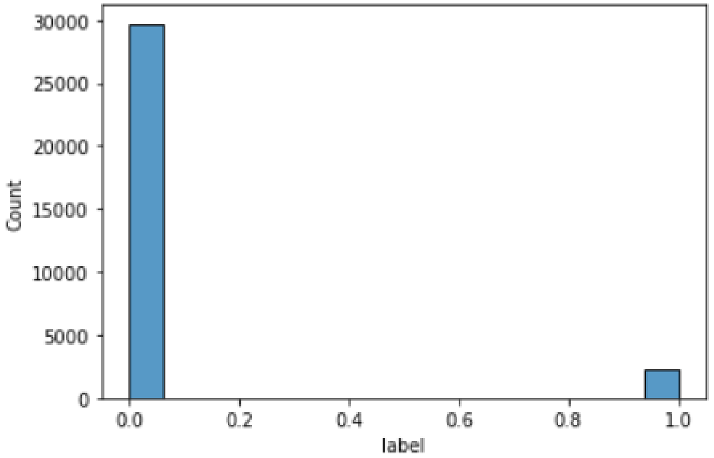
Certain words expected to influence model

- Black/white
- Racism
- supremacist
- Hate
- women

Words expected to have less/no effect

- Any word that has a general context eg, place date/time, words pertaining to nature, words elated to science/engineering etc

Data Wrangling



Only about 7% of the tweets in training data was labeled as offensive

Steps:

1. *Data was checked for null and duplicate values.*
2. *Longest tweet was sent through wordcloud which revealed several garbage characters.*
3. *The length of each tweet with respect to number of words were calculated and added to the data as a separate feature. A word tokenizer was used count the alphanumeric words in each tweet.*

	id	label	tweet	len_tweet
0	1	0	@user when a father is dysfunctional and is s...	18
1	2	0	@user @user thanks for #lyft credit i can't us...	19
2	3	0	bihday your majesty	3
3	4	0	#model i love u take with u all the time in ...	11
4	5	0	factsguide: society now #motivation	4

Length of tweet was added as a feature

WordCloud of longest tweet revealed several garbage characters

```
' @user lmfao pathetic #soit   #growup #funny #noonethere #iknowwhoitis ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x
9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x9
7ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f\x98±ð\x9f\x98±ð\x9f¤\x97ð\x9f¤\x97ð\x9f
\x98±ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð
\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82â\x80¦'
```

your	yourself	youtube	yummy	ªð	³ð	µð	¹ð	ºï	ºð	¼ð	½ð	¾ð	ó¾
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0

After adding garbage characters to stop words

wedding	week	weekend	white	wish	women	won	work	world	year	years	yes
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Data Pre-processing

- *A TfidfVectorizer was used to vectorize the words.*
- *Max features was set to 200 words; monogram and bigram words were tried; English stop words were added; also including the garbage characters. A token pattern of regex was used to find the words to vectorize.*
- *A train-test 70/30 split was done on training data*

Initial Models

- *Initial models were built with logistic regression, random forest and gradient boost and 3-hidden layer neural networks all using scikitlearn. The corresponding accuracies have been listed below:*

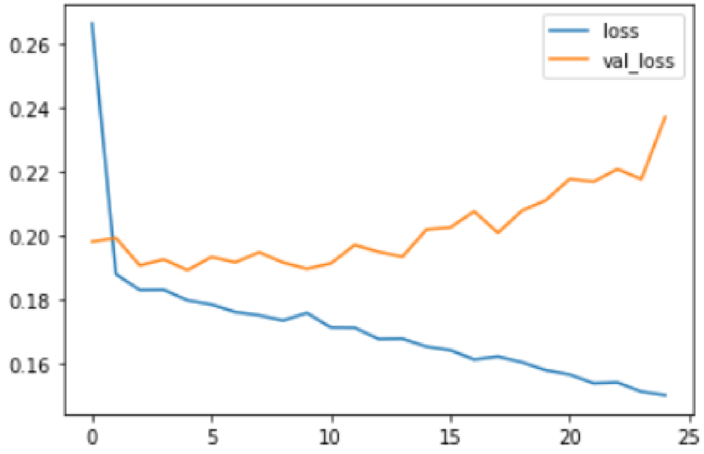
Models	Accuracy
Logistic Regression	0.932
Random forest	0.933
Gradient Boost	0.928
Neural Network	0.932

- *Based on accuracy of initial models, further optimization was carried out on Logistic Regression, Random Forest, Gradient Boost and Neural network models.*

Modelling

- Based on initial models, further Hyperparameter Tuning, Grid Search and Cross Validation was carried out using Logistic Regression, Random Forest, Decision Trees and Gradient Boost.

Model	Optimized Hyperparameter	Accuracy
Logistic Regression	C= 1	0.932
Random Forest Classifier	{'n_estimators': 250, 'min_samples_split': 10, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 227, 'bootstrap': True}	0.933
Gradient Boost Classifier	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 500}	0.933
Neural Network*	3-hidden layers, 302 nodes in each layer, Dropouts 0.9 in top and bottom, 0.5 in hidden layers, activation: ‘relu’, optimizer:’adam’; loss: ‘sparse_categorical_crossentropy’, batch-size: 256, epochs: 500, callbacks: early-stopping	0.935



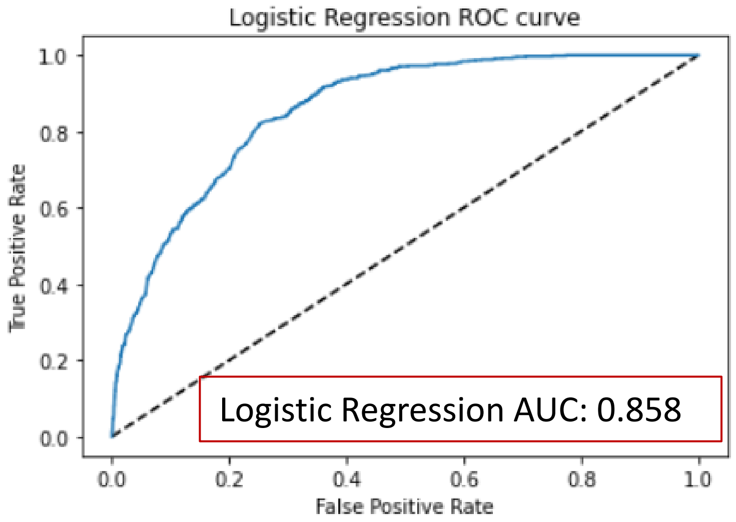
Training and validation loss with epoch for neural network model

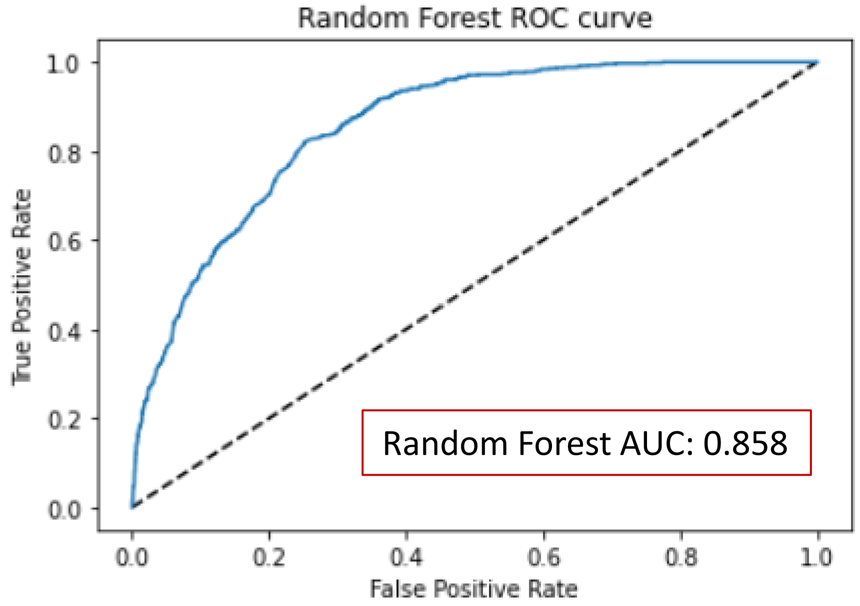
- After hyperparameter tuning, the neural network gave the highest accuracy of 0.935.

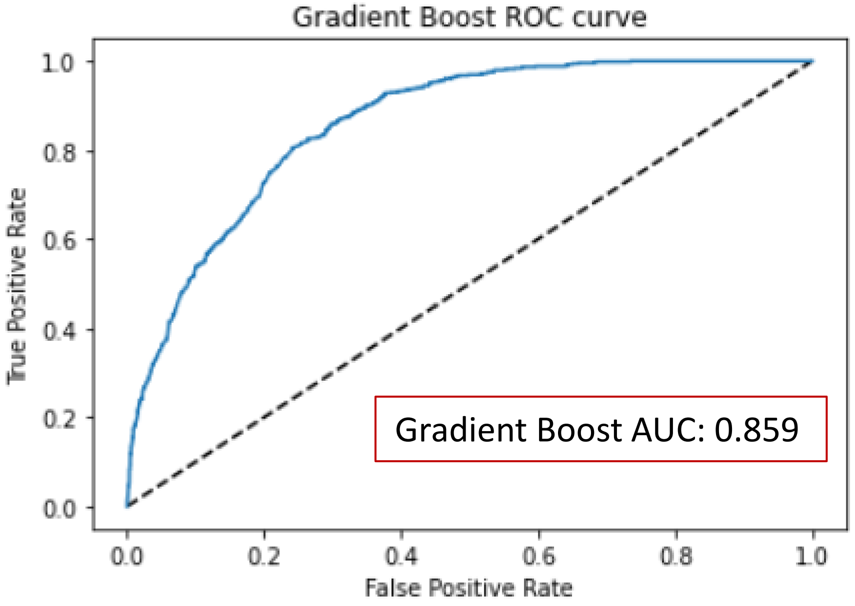
*The neural network could not be completely optimized due to time/resource constraints

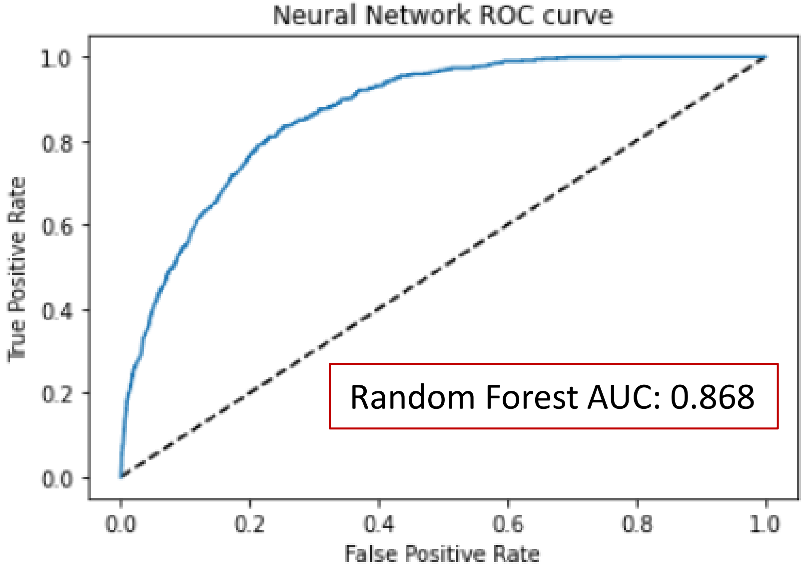
Modelling

- The ROC and AUC was calculated for each of the models.



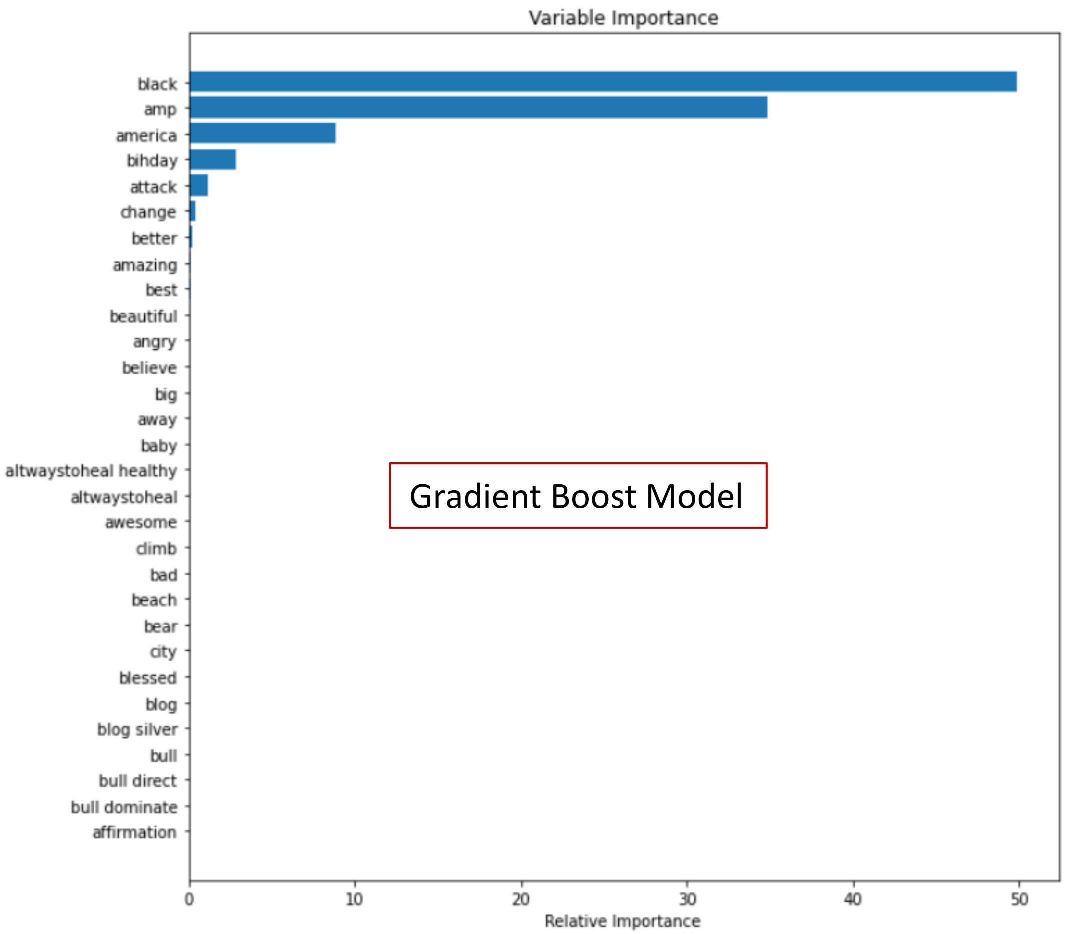


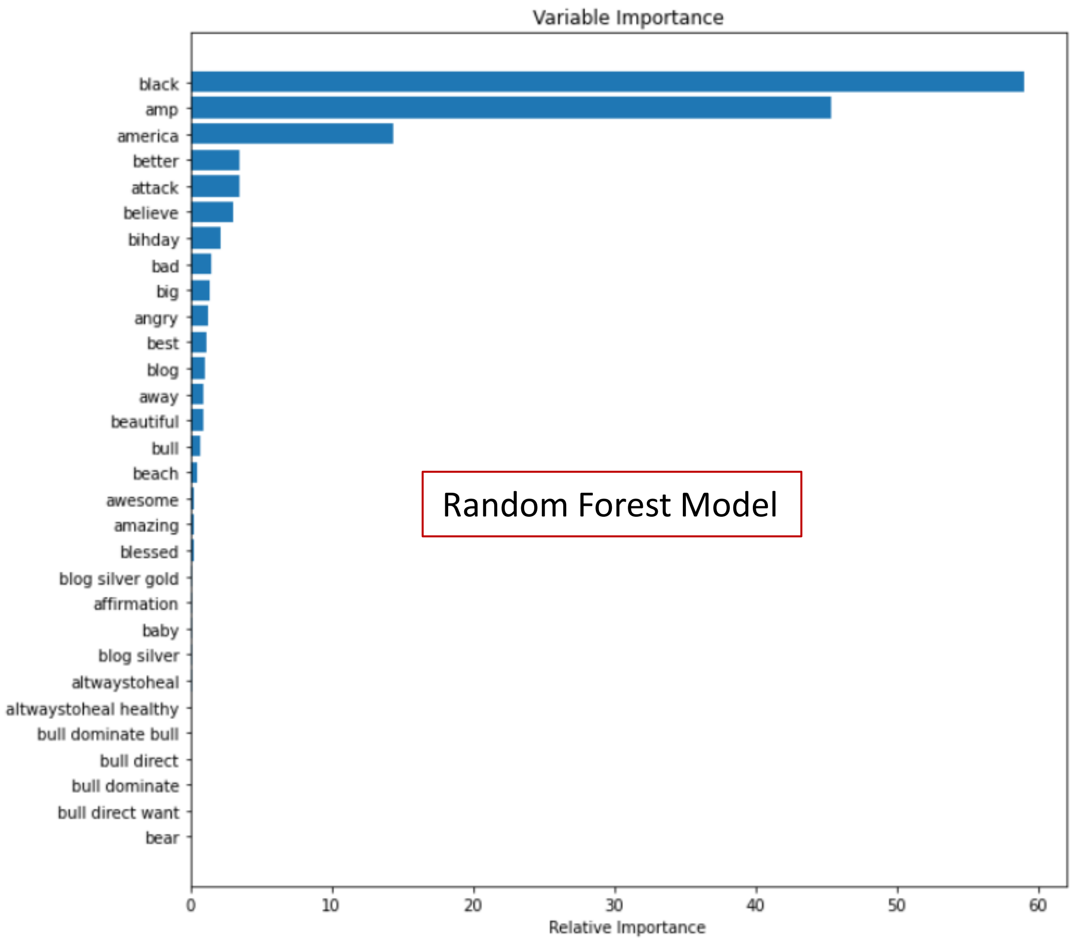




Feature Importance

- *The feature importance plots were compiled for Random Forest and Gradient Boost model.*

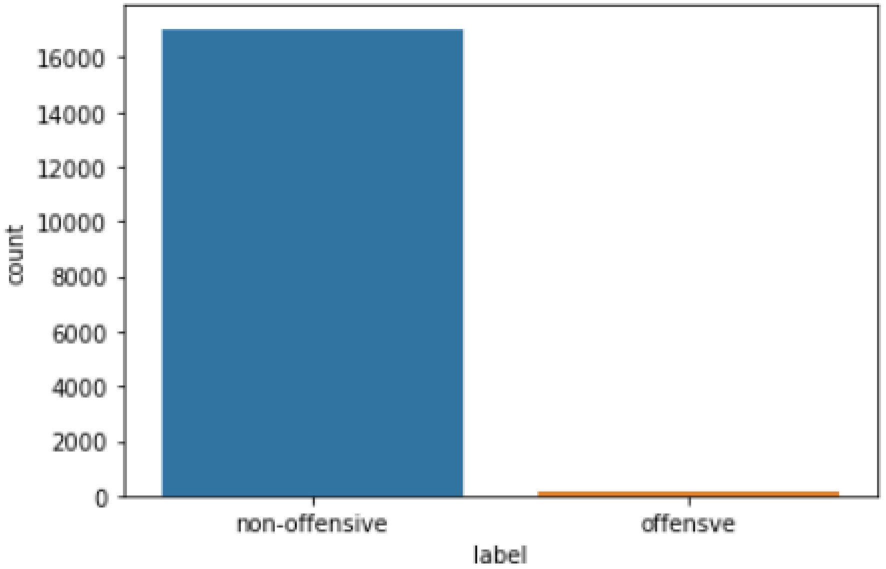




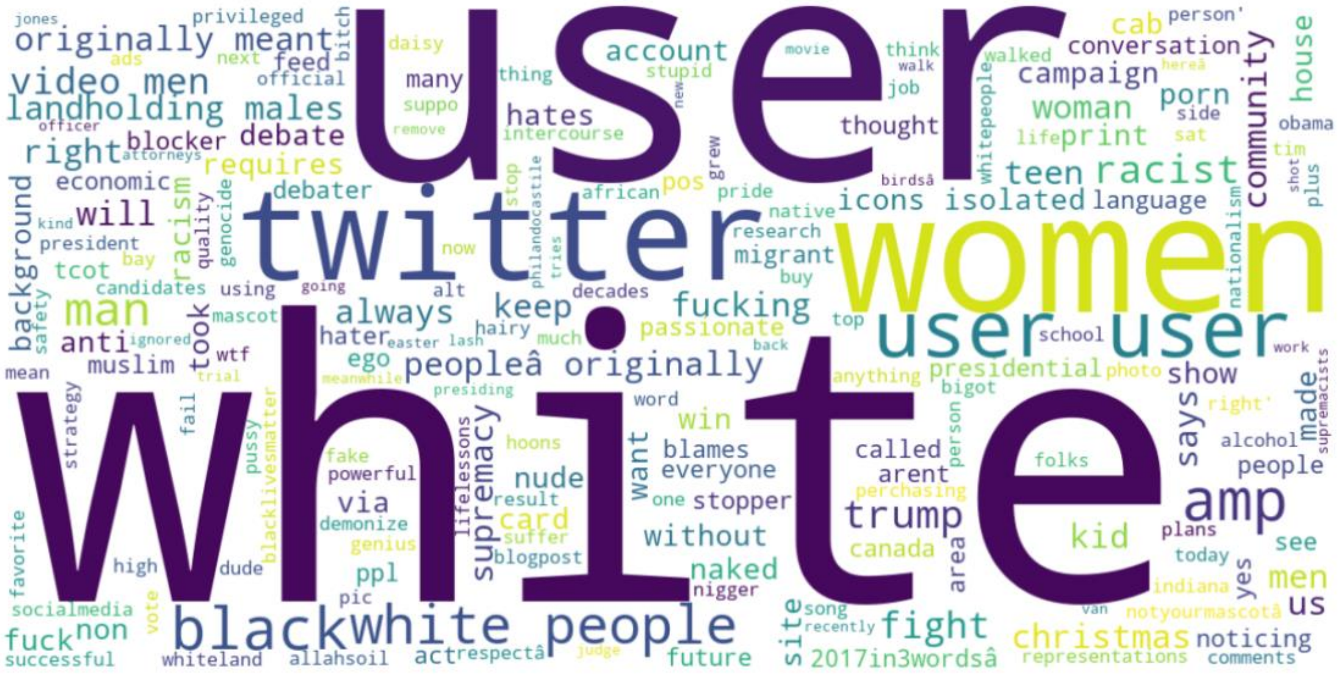
- *Both models give higher priority to words like 'black', 'America', 'attack', 'change', 'better' etc.*

Investigating Test Data

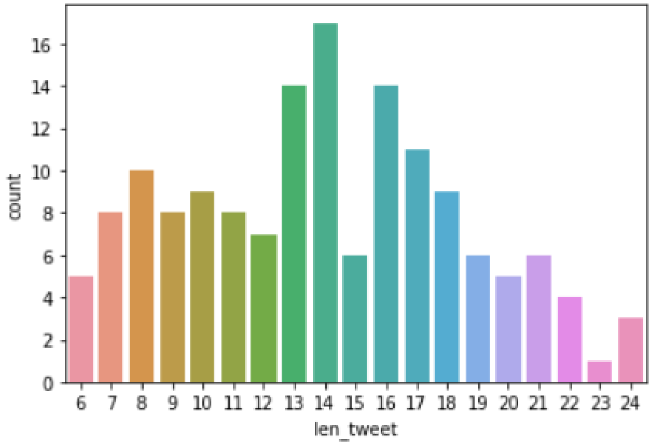
- *The semi-optimized neural network was run on test data.*
- *The model predicts 151 out of 17197 tweets to be offensive which is ~0.8%.*
- *A histogram of offensive tweets vs length was derived and a WordCloud of all 151 offensive tweets was created*



Around 0.8% of tweets were labeled as offensive on test data

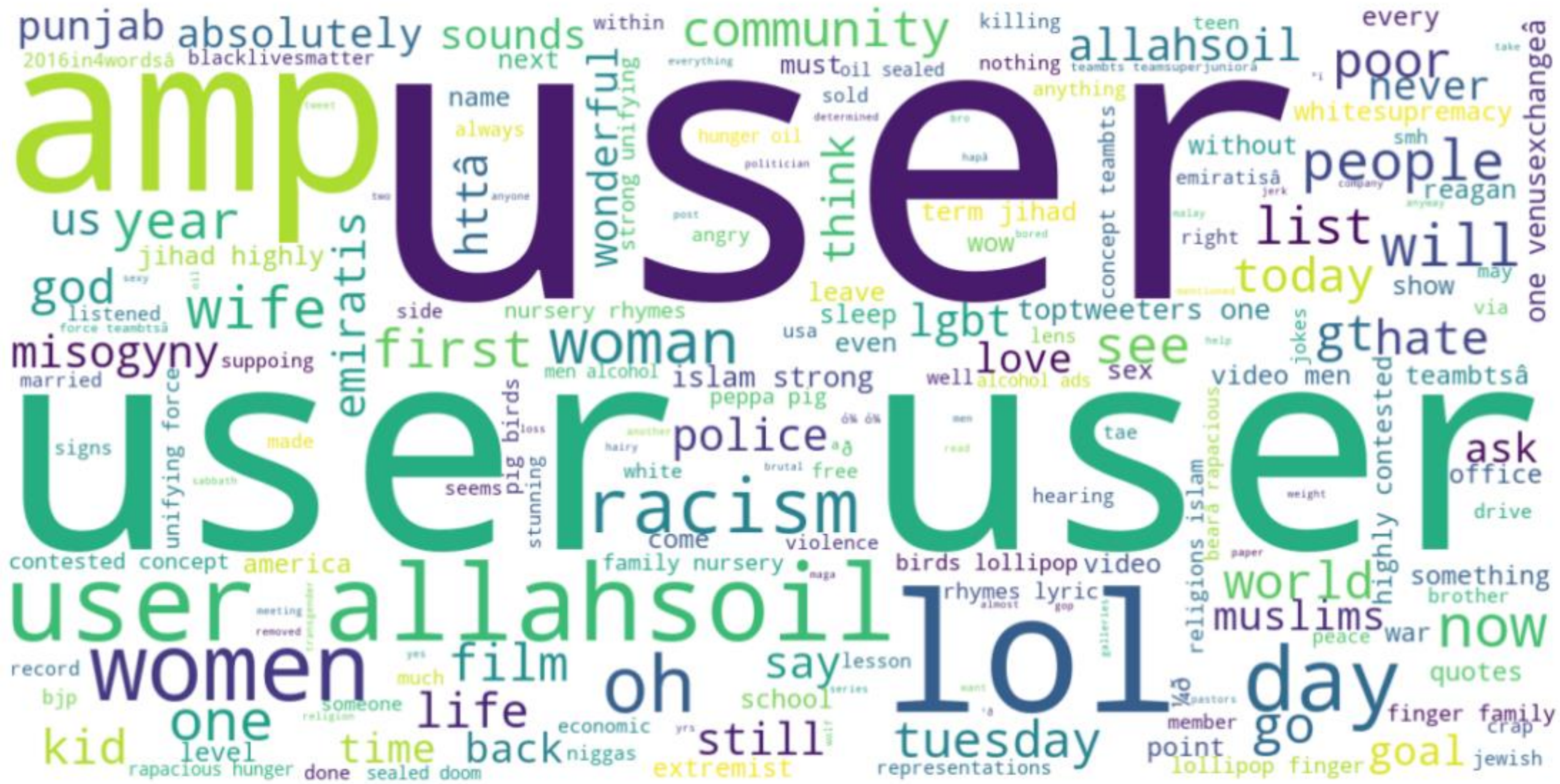


A word cloud of all offensive tweets in test data



A histogram of count vs length of all offensive tweets

Overfitting



- *With max iteration set to 800, the accuracy of the model on training data goes up to 0.947; however, when applied to test data, it is not able to correctly identify the racist tweets as clearly seen from the WordCloud. Most of the words seem to be un-offensive. This shows that increasing the max iteration too much will result in over fitting of the training data and will show large variance on test data.*