OFFENSIVE TWEET CLASSIFICATION

Ragini Rani (1610110272)
V Pratiksha Sharma (1610110415)

WHY THIS TOPIC?

Hate speech has become a major issue that is currently a hot topic in the domain of social media. Simultaneously, current proposed methods to address the issue raise concerns about censorship.

Our method utilizes publicly available machine learning models, which are tested against a hate speech corpus from Twitter.

Overview

- We will start with preprocessing and cleaning of the raw text of the tweets.
- Then we will explore the cleaned text and try to get some intuition about the context of the tweets.
- After that, we will extract numerical features from the data.
- Finally, we will build models based on the extracted features.

UNDERSTANDING THE PROBLEM STATEMENT

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.

Formally, given a training sample of tweets and labels, where label ‘1’ denotes the tweet is racist/sexist and label ‘0’ denotes the tweet is not racist/sexist, our objective was to predict the labels on the given test dataset.

TWEETS PREPROCESSING AND CLEANING

1. Removing twitter handles (@user)
2. Removing punctuations, numbers, special characters
3. Removing short words (hmm, ok, oh)
4. Tokenization
5. Stemming

id	label	tweet
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	0	bihday your majesty
4	0	#model i love u take with u all the time in urÃ°ÂŸÂ“Â±!!! Ã°ÂŸÂ˜Â™Ã°ÂŸÂ˜ÂŽÃ°ÂŸÂ‘Â„Ã°ÂŸÂ‘Â…Ã°ÂŸÂ’Â¦Ã°ÂŸÂ’Â¦Ã°ÂŸÂ’Â¦
5	0	factsguide: society now #motivation
6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyÃ¢Â€Â¦
8	0	the next school year is the year for exams.Ã°ÂŸÂ˜Â¯ can't think about that Ã°ÂŸÂ˜Â- #school #exams #hate #imagine #actorslife #revolutionschool #girl
9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers Ã¢Â€Â¦
10	0	@user @user welcome here ! i'm it's so #gr8 !

BEFORE

AFTER

	id	label	tweet	tidy_tweet
0	1	0.0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunctional selfish drags kids into dysfunction #run
1	2	0.0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #lyft credit cause they offer wheelchair vans #disapointed #getthanked
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ ð□□¦ð□□¦ð□□¦	#model love take with time
4	5	0.0	factsguide: society now #motivation	factsguide society #motivation

1. Understanding the common words used in the tweets
2. Visualising words in non racist/sexist tweets
3. Visualising words in racist/sexist tweets
4. Understanding the impact of hashtags on tweets sentiment

UNDERSTANDING AND VISUALISATION FROM TWEETS

1. Common words: wordcloud

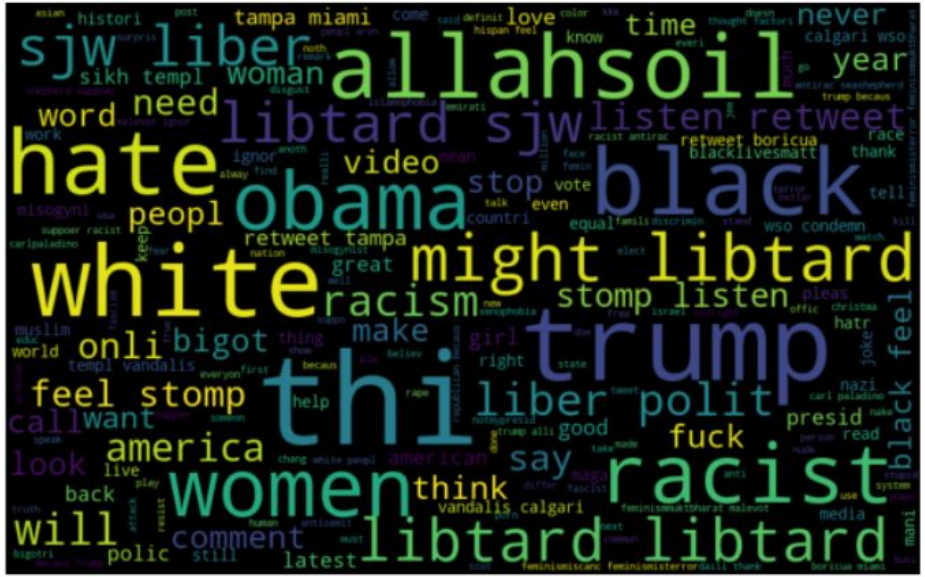
In [17]:
```
all_words = ' '.join([text for text in combi['tidy_tweet']])
from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(all_words)

plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

2. NON RACIST/SEXIST TWEETS: WORDCLOUD

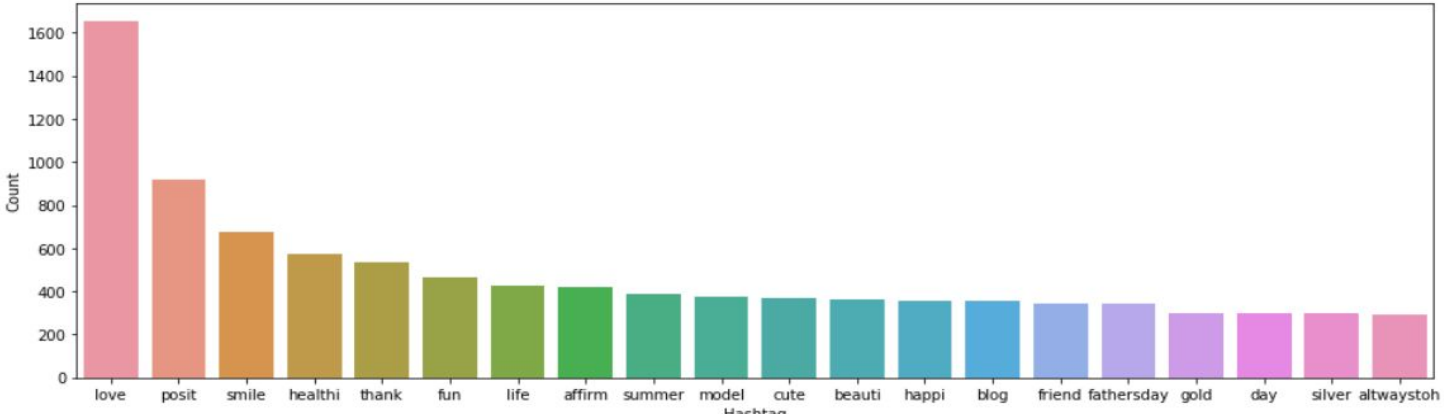
In [22]:
```
normal_words =' '.join([text for text in combi['tidy_tweet'][combi['label'] == 0]])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

3. RACIST/SEXIST TWEETS: WORDCLOUD

In [23]:
```
negative_words = ' '.join([text for text in combi['tidy_tweet'][combi['label'] == 1]])
wordcloud = WordCloud(width=800, height=500,
random_state=21, max_font_size=110).generate(negative_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

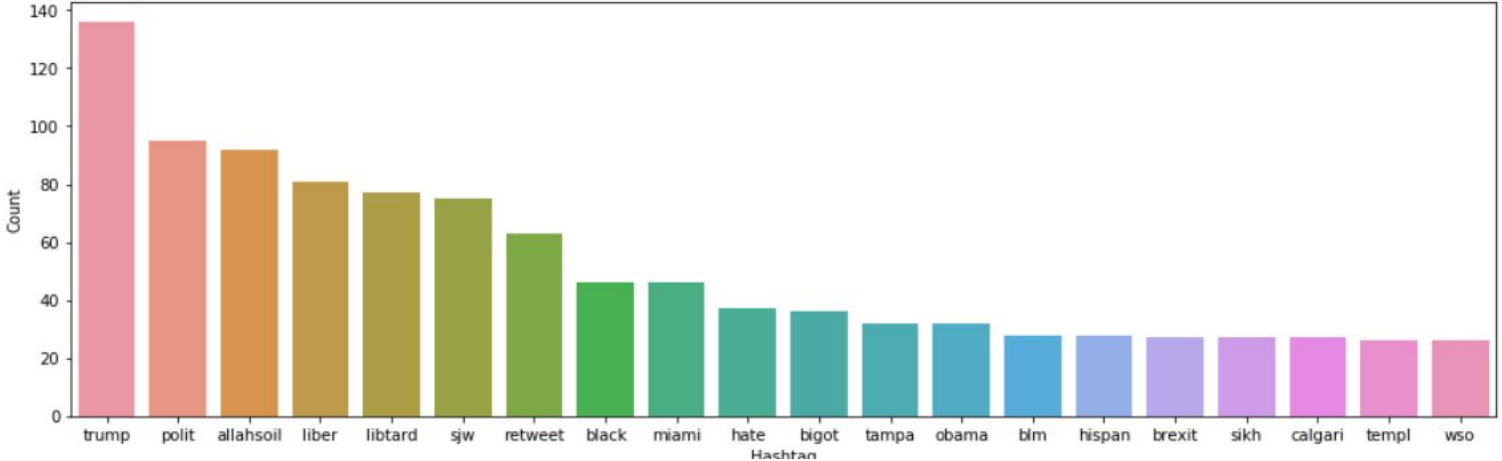


4. HASHTAGS ASSOCIATED WITH TWEETS: GRAPH



POSITIVE

NEGATIVE



EXTRACTING FEATURES FROM CLEANED TWEETS

1. Bag of Words Features
2. Tf-idf Features

1. Bag of Words Features

Bag-of-Words is a method to represent text into numerical features. Consider a corpus (a collection of texts) called C of D documents {d1,d2…..dD} and N unique tokens extracted out of the corpus C. The N tokens (words) will form a list, and the size of the bag-of-words matrix M will be given by D X N. Each row in the matrix M contains the frequency of tokens in document D(i).

```
In [30]: bow_vectorizer = CountVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')
         bow = bow_vectorizer.fit_transform(combi['tidy_tweet'])
         bow.shape

Out[30]: (49159, 1000)
```

2. TF-IDF FEATURES

TF-IDF works by penalizing the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus but appear in good numbers in few documents.

```
In [31]: tfidf_vectorizer = TfidfVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')
         tfidf = tfidf_vectorizer.fit_transform(combi['tidy_tweet'])
         tfidf.shape

Out[31]: (49159, 1000)
```

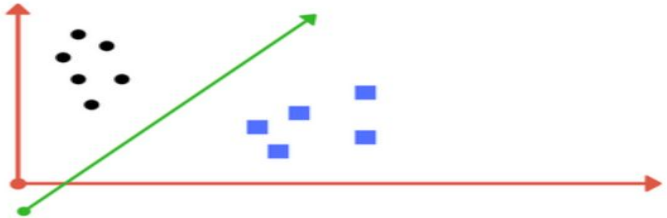
MODEL BUILDING

1. Support Vector Machine
2. Logistic Regression
3. RandomForest

1. SUPPORT VECTOR MACHINES

Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Image A: Draw a line that separates black circles and blue squares.



s Image B: Sample cut to divide into two classes.

F1 SCORE (training data):

A. BAG OF WORDS FEATURES- 0.503
B. TF-IDF FEATURES- 0.510

2. LOGISTIC REGRESSION

It predicts the probability of occurrence of an event by fitting data to a logit function. The following equation is used in Logistic Regression:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

F1 SCORE (training data):

A. BAG OF WORDS FEATURES- 0.530
B. TF-IDF FEATURES- 0.544

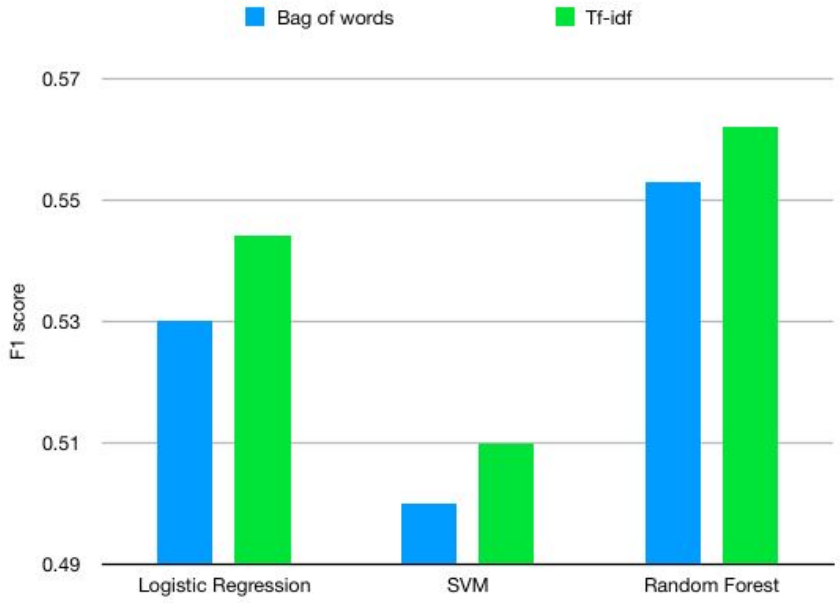
3. RANDOM FOREST

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as follows, where the final model g is the sum of simple base models fi.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

F1 SCORE (training data):

A. BAG OF WORDS FEATURES- 0.552
B. TF-IDF FEATURES- 0.562


Bag of words
Tf-idf
0.57
0.55
0.53
0.51
0.49
F1 score
Logistic Regression
SVM
Random Forest

ANALYSIS


professional twiter name
@sarahjeong
Are white people genetically predisposed to burn faster in the sun, thus logically being only fit to live underground like groveling goblins
9:23 PM - 23 Dec 2014
47 Retweets 106 Likes
seal Papachristou
@ papaxristoutj
Follow
With so many Africans in Greece .. At least the mosquitoes of West Nile .. will eat homemade food!
Reply Retweet Favorite
101
46
Follow
[redacted]
First thing my mom says this morning: did you hear the bad news? The monkey is staying for another 4 years...
#WeHateYouObama
Reply Retweet Favorite

THANK YOU