Hate Speech Detection

Exploratory Data Analysis of Tweets Dataset

Group Name: VioletGo
xw3080@nyu.edu

2023 Jul 21

Xinyu Wang

Problem Description

I'm developing an NLP-based classifier to detect offensive or hate speech in tweets, aiming to foster healthier online dialogues. It identifies harmful content, promoting respectful community interactions on Twitter. The precision, derived from advanced NLP and machine learning techniques, differentiates between harmless banter and genuinely offensive language. This project represents a convergence of technological advancement and social responsibility, striving to protect users from hate speech and foster a more positive online ecosystem.

Data Description

https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv

We're working with a collection of tweets, each associated with a unique ID and a label. There are 31,962 such entries in our dataset.

Three key elements:

1. ID: A unique identifier for each tweet.
2. Label: A marker indicating if the tweet is 'hate speech' (1) or not (0).
3. Tweet: The actual text content of the tweet.

Our aim is to leverage this data to train a model, which will later predict the 'Label' for unseen tweets in a separate test set of 17,197 entries. The test set contains 'ID' and 'Tweet' but not 'Label' - that's what we'll predict.

```
In [98]: df_train.head()
```

Out[98]:

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

```
In [99]: df_test.head()
```

Out[99]:

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedic...
1	31964	@user #white #supremacists want everyone to s...
2	31965	safe ways to heal your #acne!! #altwaystohe...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd #bihday to my amazing, hilarious #nephew...

Data Cleaning(regex,nltk)

- Lowercasing: Standardizing case.
- Remove Noise: Strip numbers, punctuation, special chars.
- Tokenization: Break down text into words.
- Stopword Removal: Remove common, non-informative words.
- Stemming: Reduce words to their root form.
- Rejoining: Combine processed words into a string.

Vectorization

We utilized three different methods to extract meaningful features from our preprocessed tweet data:

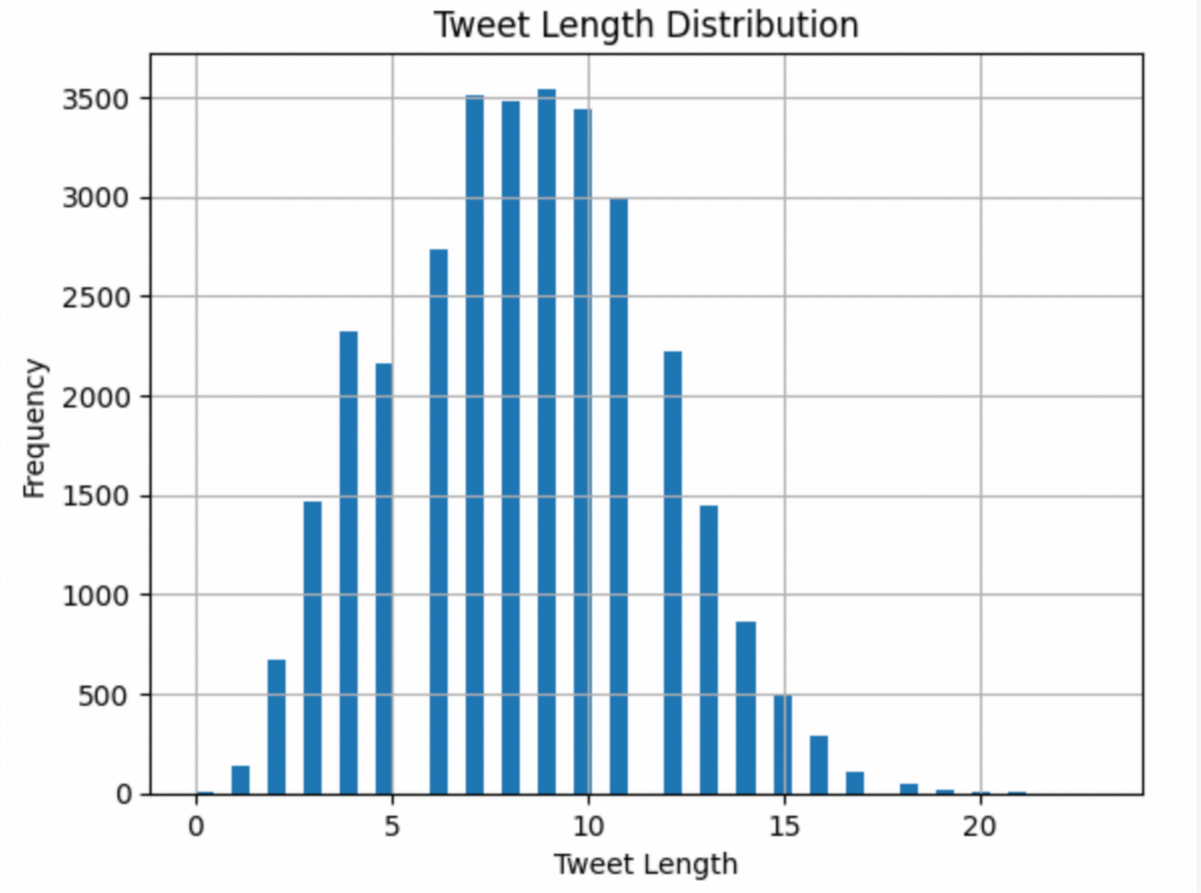
- TF-IDF: A numerical statistic that reflects how important a word is to a document in a collection or corpus. We created 500 features with this method.
- Word2Vec: A popular technique for natural language processing that uses a neural network to learn word associations. We also generated 500 features using this approach.
- BERT: An advanced method developed by Google that uses transformer architecture to understand the context of words in sentences. It produced 768 features.

EDA:Length of Tweets

The histogram showcases the distribution of tweet lengths, following a bell curve, albeit with a long tail(skewed right), indicating that our tweets vary considerably in length.

Key points to note:

1. Mode: Most tweets fall within a length of 2-15 words, clustering near the middle of the distribution.
2. Long Tail: The tail of the distribution extends further to the right. This signifies that while most tweets are quite succinct, there are also several longer comments and replies, creating a diverse range of tweet lengths in our dataset.

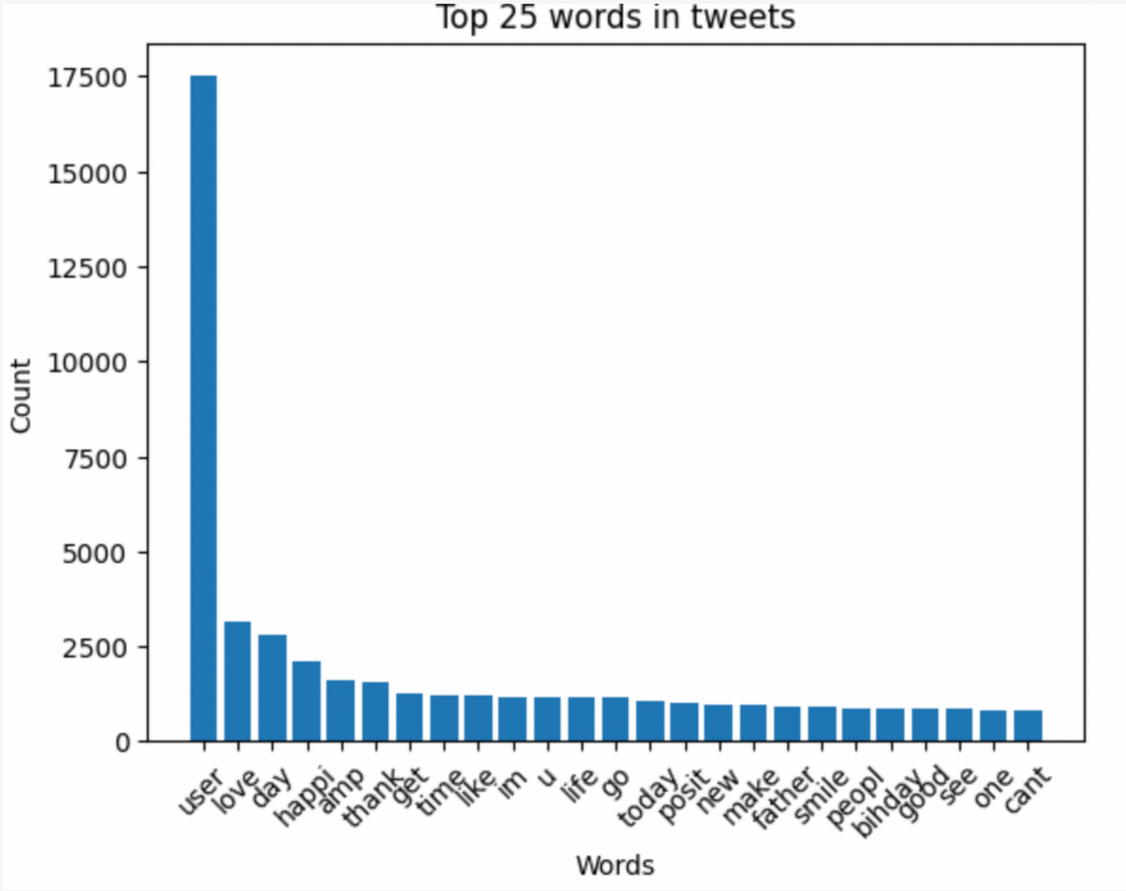


EDA: Most Common Words

The bar chart displays the 25 most frequently occurring words in our dataset. These words provide insight into the most common themes present in our tweets.

Key points to note:

1. 'user' Dominance: The term 'user', appearing almost 17,500 times, stands out as the most frequent term. This high frequency is largely due to the common practice of mentioning other users in tweets ('@user').
2. Positive Tone: Words like 'love', 'happy', 'thank', and 'life' appear frequently, suggesting a generally positive or neutral tone in the majority of tweets.
3. Common Words: Words related to everyday experiences (e.g., 'day', 'time', 'get', 'go', 'today') also feature prominently, indicating the routine nature of many tweets.

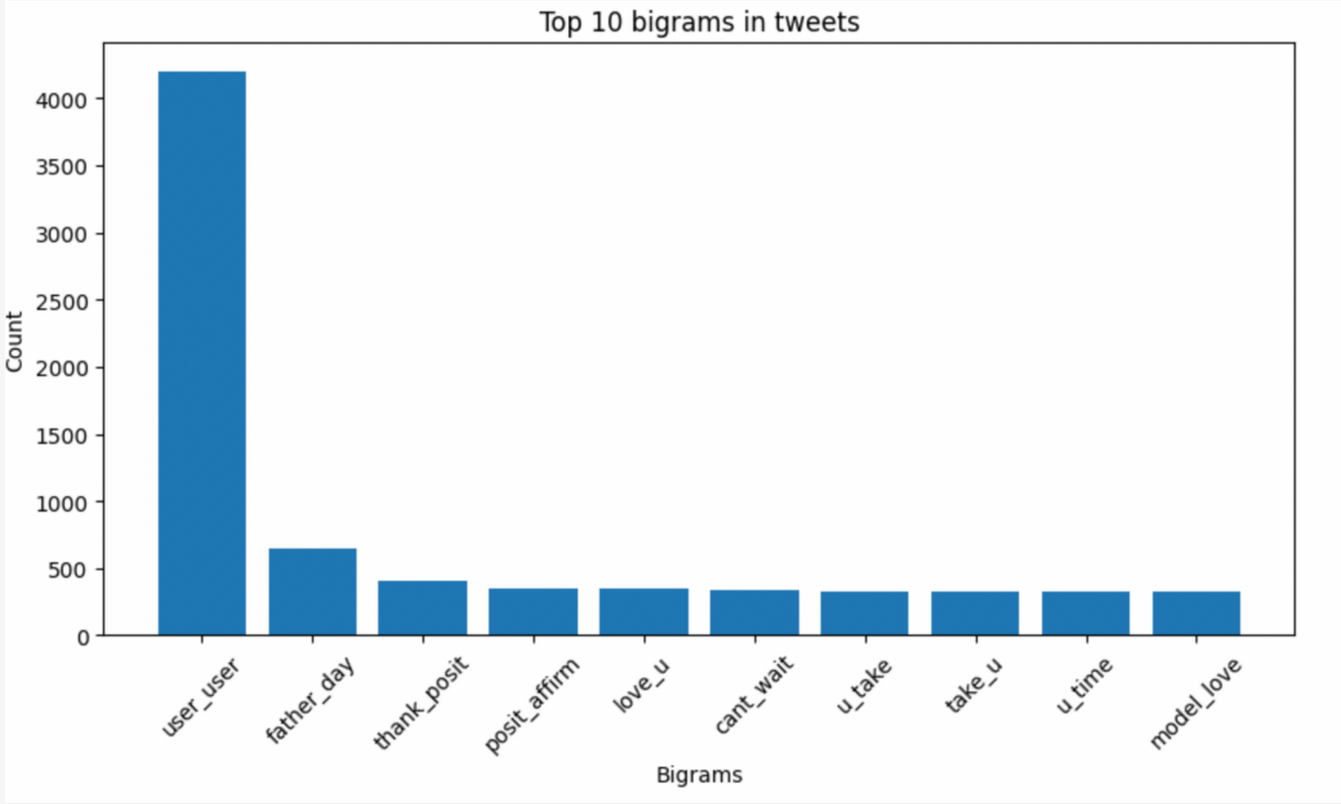


EDA: Most Common Bigrams

Bigrams are pairs of consecutive written units such as letters, syllables, or words. They provide a more nuanced understanding of the context and can reveal recurring themes or common phrases in our tweets.

Key insights from the top 10 bigrams:

1. **'user_user' Dominance**: Similar to the individual word frequency, 'user_user' (likely '@user @user') is the most frequent bigram with around 4,000 occurrences. This bigram underlines the interactive nature of Twitter, with users frequently communicating with each other.
2. **Special Occasions & Sentiments**: Phrases like 'father_day' and 'cant_wait' suggest that special occasions and anticipatory sentiments form part of the common discussions.
3. **Positive Affirmations**: Bigrams like 'thank_posit', 'posit_affirm', and 'love_u' indicate a recurring theme of positivity and gratitude in the tweets.

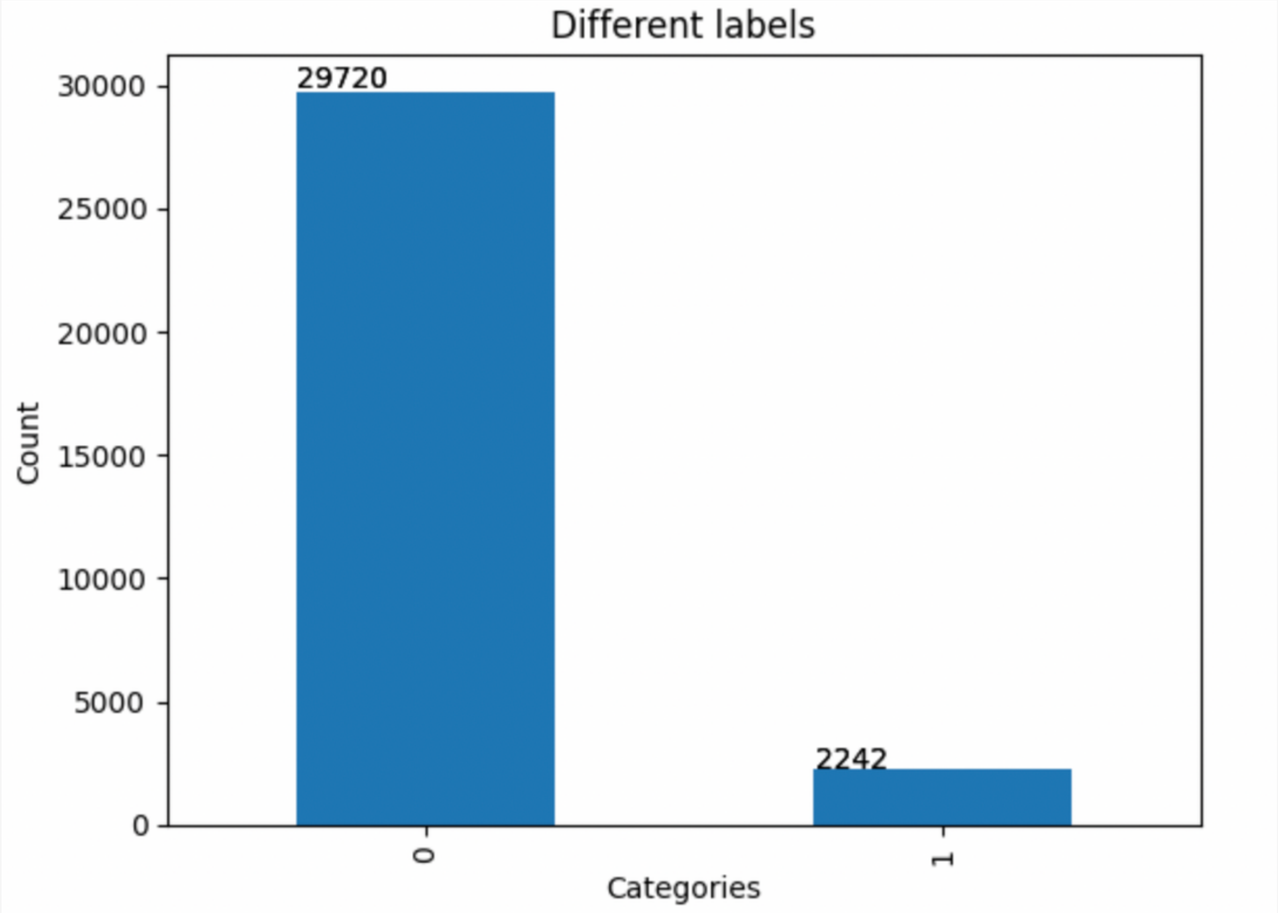


EDA: Tweet Label Distribution

Understanding the distribution of labels in our dataset is crucial as it impacts how we approach model training and performance evaluation.

The bar graph showcases a significant imbalance between the two classes:

1. Label 0 (Non-hate speech): The majority of tweets fall into this category, indicating a generally positive or neutral tone in the discussions.
2. Label 1 (Hate speech): Comprising about 7.014% (2242 tweets) of the dataset, this category is considerably smaller. It's important to note that due to this imbalance, we may need to use specific techniques to handle imbalanced classes in our predictive modeling.

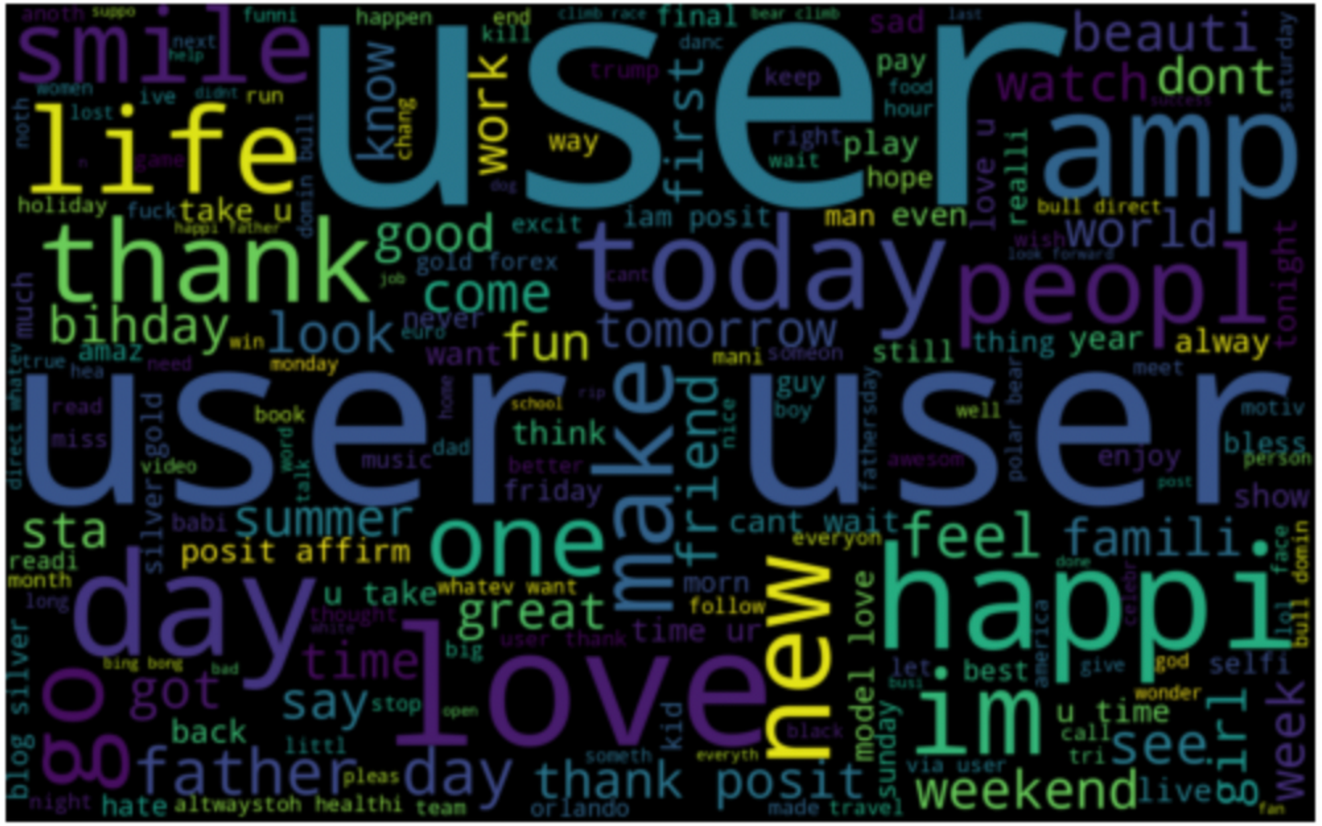


EDA: Word Clouds of all tweets

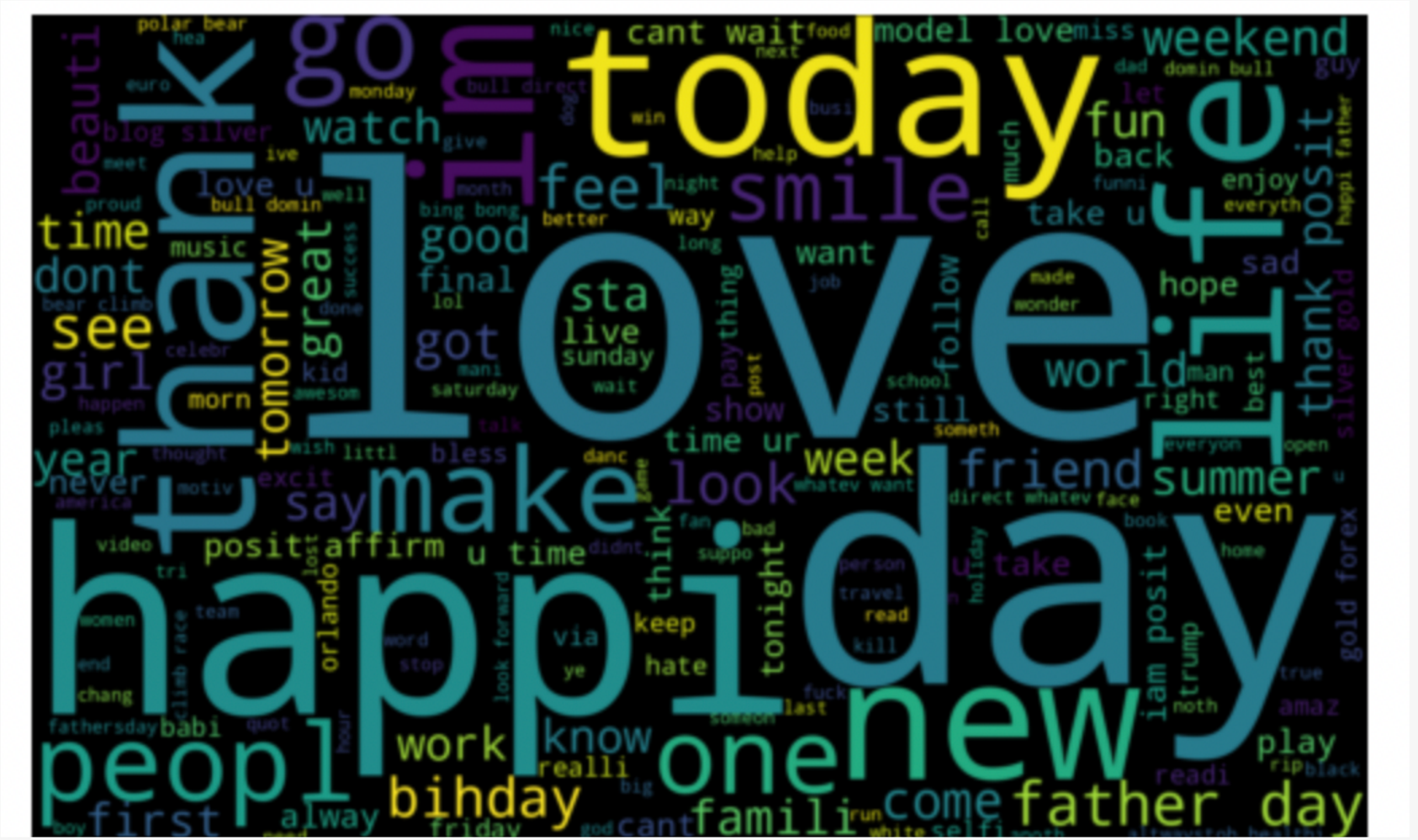
A word cloud is an excellent tool for visualizing the most frequent words in our dataset. The bigger the word in the cloud, the more frequently it appears in our data.

The word cloud for all tweets (including both non-hate and hate speech) is largely dominated by words such as 'user', 'today', 'thank', etc. This gives us a quick and easy to understand overview of the most common themes or topics in the entire dataset.

Please note: The word 'user' likely appears frequently due to the common practice of tagging other users in tweets, rather than indicating a specific topic or theme.



EDA: Word Cloud - all tweets

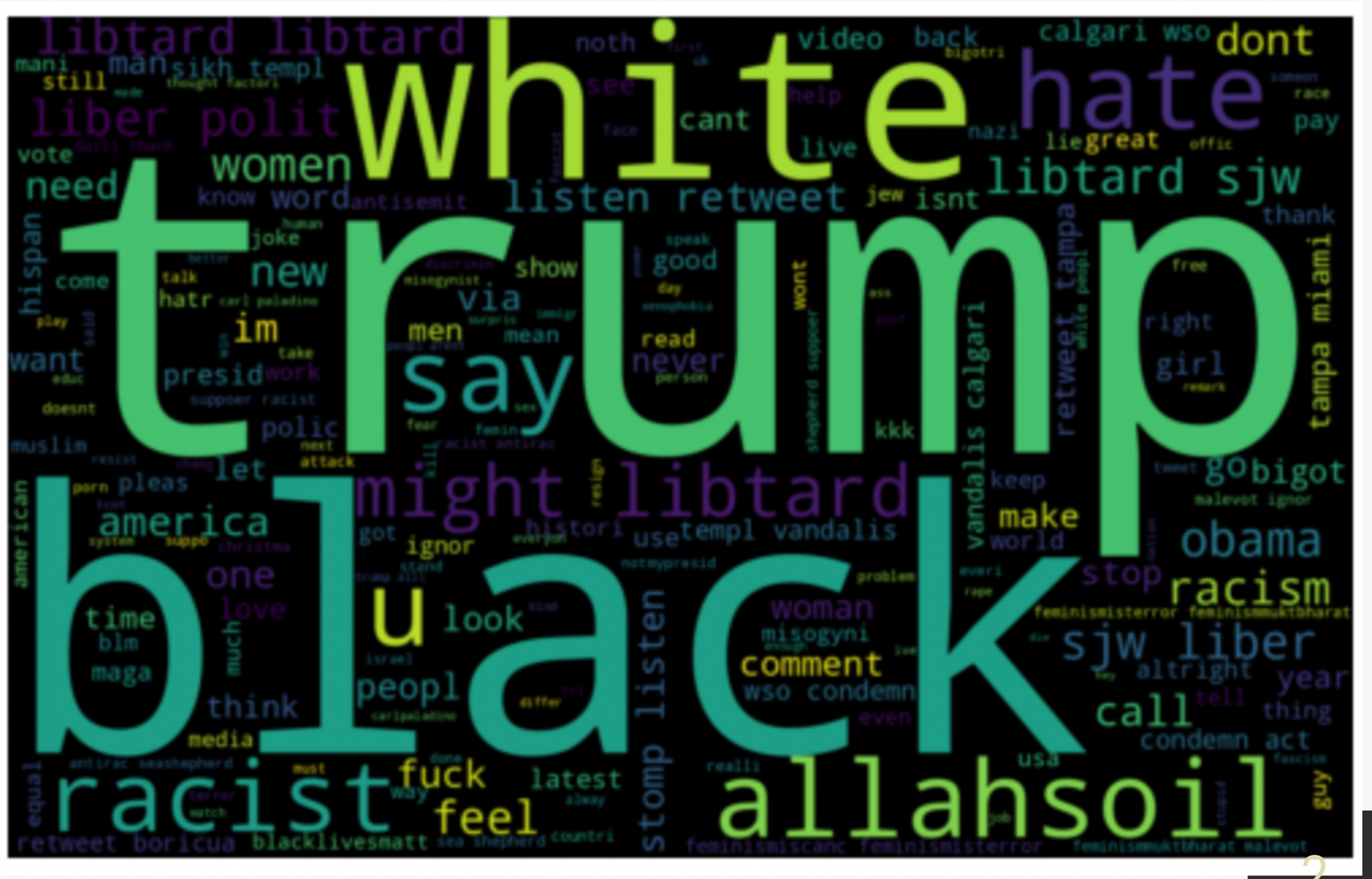


EDA: Word Cloud - hate speech

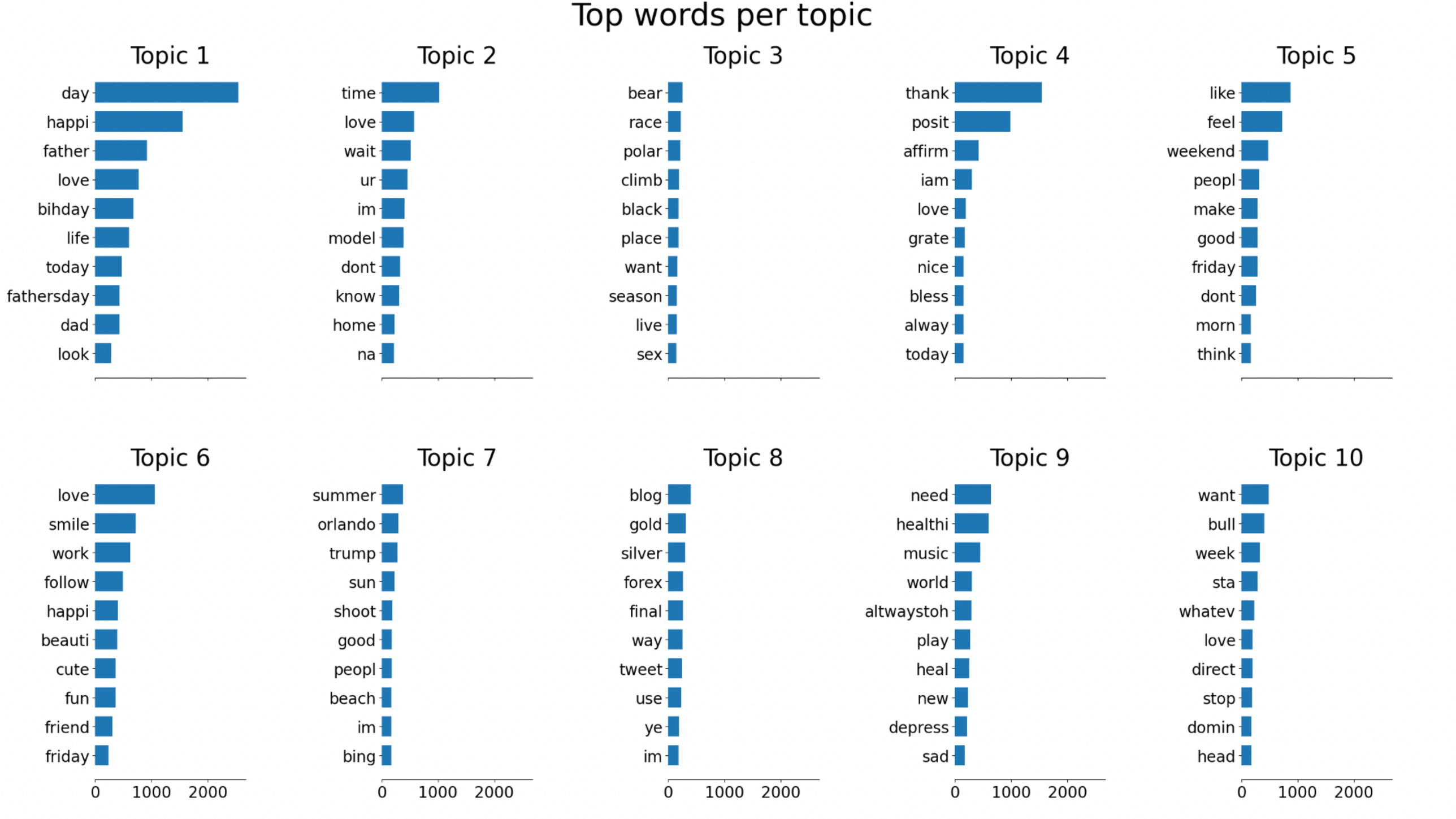
Our word cloud highlights key themes in hate speech:

- Politically charged terms: 'Trump'
- Racial identifiers: 'black', 'white'
- Direct indicators: 'racist', 'hate'

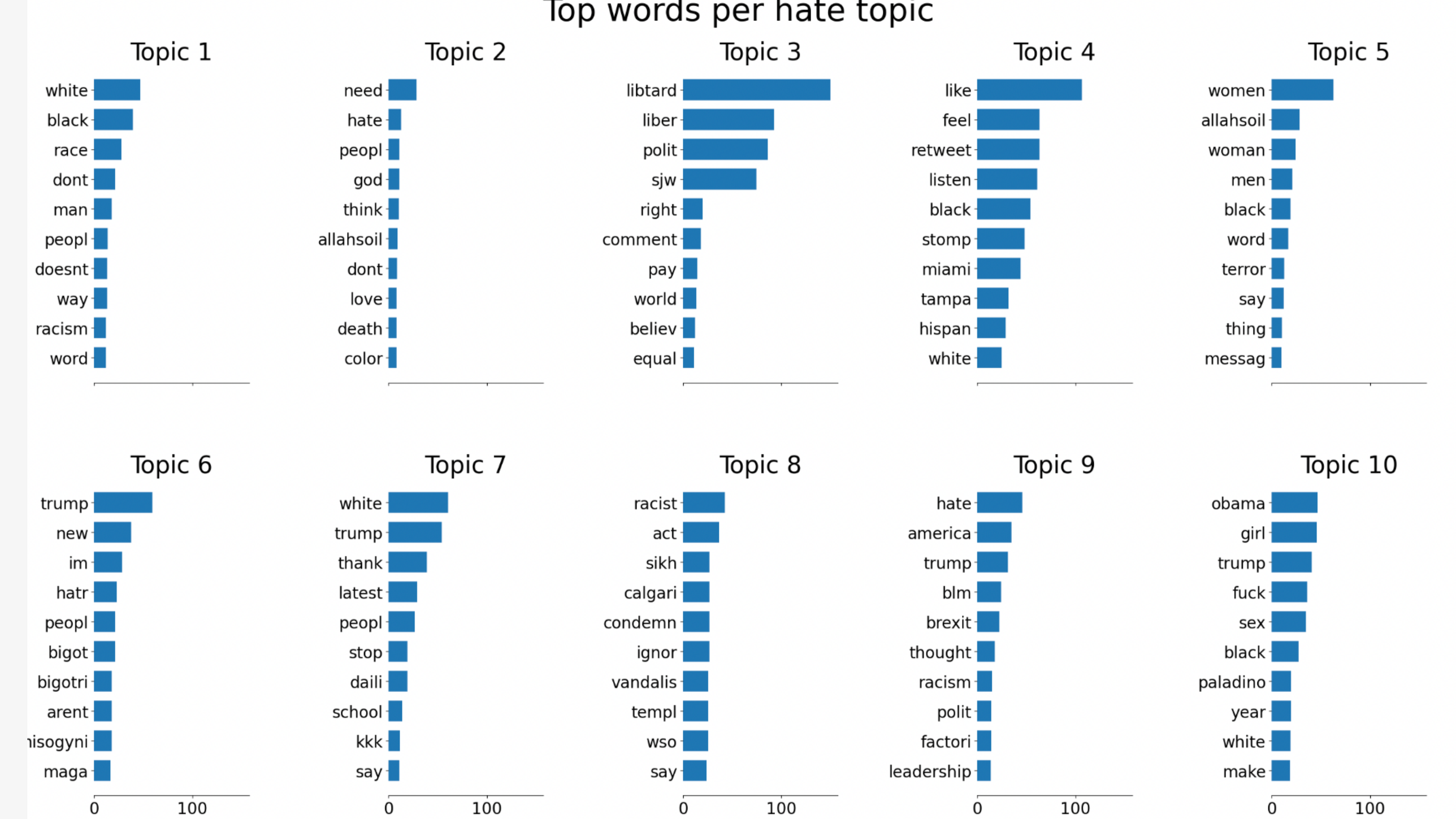
Understanding these common terms provides insight into hate speech patterns in our dataset.



EDA: Topics of all tweets


Top words per topic
Topic 1
day
happi
father
love
bihday
life
today
fathersday
dad
look
Topic 2
time
love
wait
ur
im
model
dont
know
home
na
Topic 3
bear
race
polar
climb
black
place
want
season
live
sex
Topic 4
thank
posit
affirm
iam
love
grate
nice
bless
alway
today
Topic 5
like
feel
weekend
peopl
make
good
friday
dont
morn
think
Topic 6
love
smile
work
follow
happi
beauti
cute
fun
friend
friday
0
1000
2000
Topic 7
summer
orlando
trump
sun
shoot
good
peopl
beach
im
bing
0
1000
2000
Topic 8
blog
gold
silver
forex
final
way
tweet
use
ye
im
0
1000
2000
Topic 9
need
healthi
music
world
altwaystoh
play
heal
new
depress
sad
0
1000
2000
Topic 10
want
bull
week
sta
whatev
love
direct
stop
domin
head
0
1000
2000

EDA: Topics of hate speech


Top words per hate topic
Topic 1
white
black
race
dont
man
peopl
doesnt
way
racism
word
Topic 2
need
hate
peopl
god
think
allahsoil
dont
love
death
color
Topic 3
libtard
liber
polit
sjw
right
comment
pay
world
believ
equal
Topic 4
like
feel
retweet
listen
black
stomp
miami
tampa
hispan
white
Topic 5
women
allahsoil
woman
men
black
word
terror
say
thing
messag
Topic 6
trump
new
im
hatr
peopl
bigot
bigotri
arent
misogyni
maga
0
100
Topic 7
white
trump
thank
latest
peopl
stop
daili
school
kkk
say
0
100
Topic 8
racist
act
sikh
calgari
condemn
ignor
vandalis
templ
wso
say
0
100
Topic 9
hate
america
trump
blm
brexit
thought
racism
polit
factori
leadership
0
100
Topic 10
obama
girl
trump
fuck
sex
black
paladino
year
white
make
0
100

Recommended Models

- We suggest the following models for hate speech detection:
- Logistic Regression: A statistical model that's effective for binary classification tasks.
- Gradient Boosting Machine: An ensemble machine learning model that often performs well on a variety of datasets.
- Neural Networks: Capable of learning complex patterns and relationships, especially beneficial for text data.
- BERT (Bidirectional Encoder Representations from Transformers): A state-of-the-art NLP model, highly effective for text classification tasks like this.

Thank you for listening