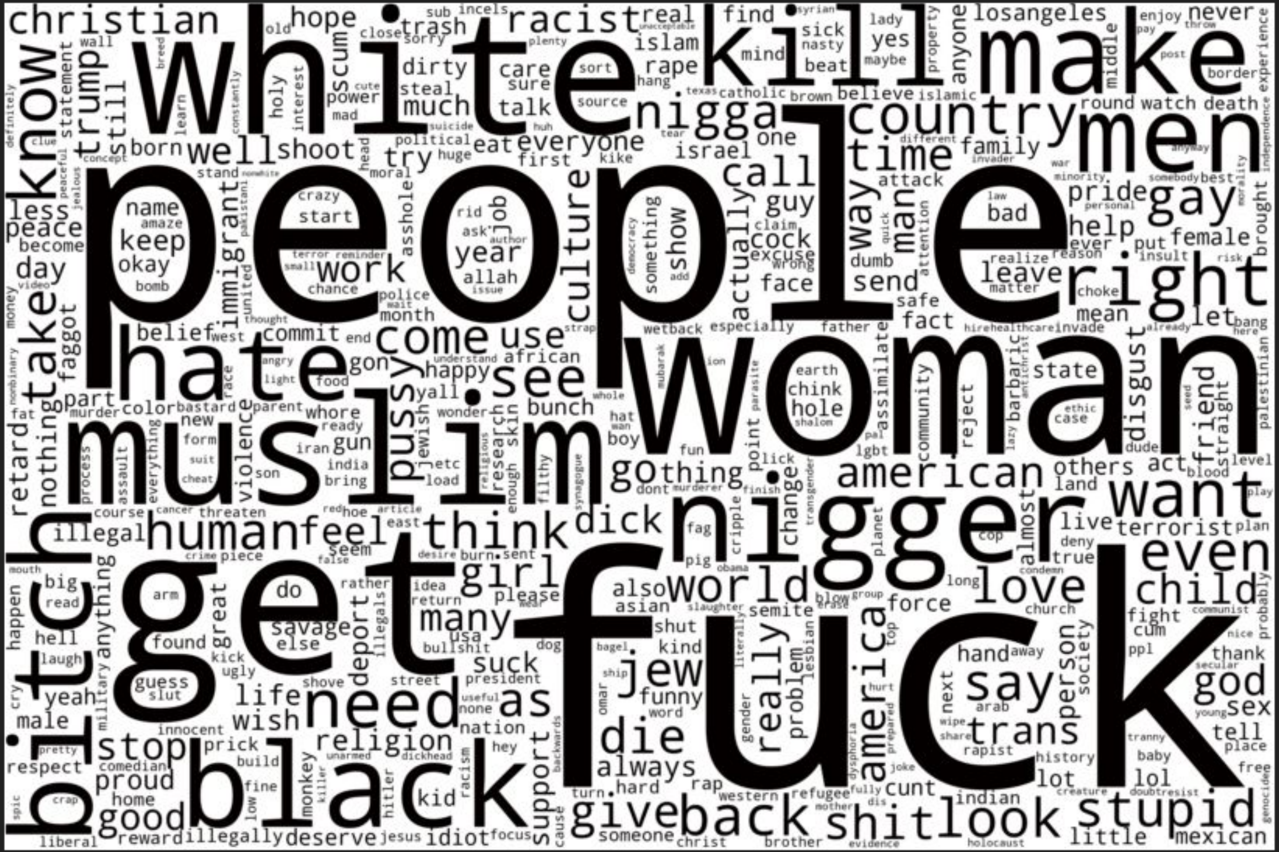
HATE SPEECH DETECTION

MACHINE LEARNING PROJECT

BY
ABHISHEK SINGH YADAV
(20HCS4103)

The problem:


people
woman
fuck
white
kill
make
get
muslim
black
nigger
hate
bitch
know
men
country
nigga
right
want
even
child
love
world
give
back
shit
look
stupid
say
trans
american
think
human
feel
need
dick
die
jew
america
really
god
christian
racist
hope
trash
islam
rape
sick
yes
losangeles
never
time
call
culture
pussy
take
come
see
use
work
year
allah
immigrant
well
shoot
try
born
trump
still
talk
care
sure
steal
much
dirty
scum
power
everyone
israel
family
gay
pride
help
female
put
leave
face
send
fact
way
man
guy
cock
show
something
actually
dumb
bad
state
disgust
friend
straight
others
act
terrorist
live
girl
many
also
asian
semite
force
hand
person
society
history
lot
lol
tell
sex
place
free
mexican
little
cunt
indian
rapist
arab
religion
stop
proud
good
home
respect
male
yeah
guess
slut
found
savage
else
deport
kid
idiot
jesus
deserve
illegally
reward
liberal
support
always
hard
rap
western
refugee
funny
word
nation
suck
president
happy
yall
bunch
skin
research
jewish
violence
gun
india
bring
whore
ready
thing
point
change
problem
lesbian
community
reject
barbaric
assimilate
earth
chink
hole
name
keep
okay
day
less
peace
become
faggot
retard
nothing
fat
part
murder
color
bastard
new
form
illegal
course
threaten
great
anything
happen
hell
laugh
big
read
arm
do
wish
life
prick
build
fine
monkey
hitler
racism
cripple
fag
pig
shut
kind
death
watch
round
border
middle
anyone
property
maybe
beat
nasty
mind
find
real
believe
islamic
catholic
brown
one
first
eat
kike
political
huge
moral
stand
crazy
start
job
rid
ask
author
asshole
police
month
belief
west
commit
end
gon
african
food
light
angry
race
parent
iran
son
load
filthy
enough
boy
hat
dont
finish
synagogue
thank
sub
incels
old
close
sorry
wall
statement
holy
interest
learn
mad
suicide
definitely
concept
peaceful
jealous
amaze
bomb
video
money
chance
issue
terror
reminder
small
united
pakistani
nonbinary
process
assault
everything
military
cry
pretty
comedian
crap
spic
innocent
unarmed
killer
low
ugly
kick
shove
street
useful
none
hey
dickhead
burn
sent
desire
idea
return
rather
false
crime
piece
mouth
usa
bullshit
dog
bagel
ship
omar
gender
cause
backwards
turn
someone
christ
brother
evidence
holocaust
doubtresist
creature
baby
tranny
young
secular
nice
communist
fight
cum
church
condemn
true
deny
cop
almost
planet
transgender
lgbt
fun
lazy
case
ethic
dude
blood
level
play
plan
probably
genocide
palestinian
bang
here
let
mean
invade
already
hirehealthcare
father
wetback
especially
choke
matter
realize
reason
insult
risk
brought
best
morality
somebody
minority
war
invader
different
anyway
independence
experience
post
pay
throw
enjoy
lady
syrian
nasty
unacceptable
plenty
hang
texas
source
tear
huh
sort
cute
democracy
wrong
excuse
claim
attention
quick
law
personal
safe
matter
antichrist
whole
mubarak
parasite
shalom
pal
lick
murderer
slaughter
literally
blow
erase
group
top
long
next
hurt
wipe
share
prepared
joke
dysphoria
fully
dis
mother
please
sent
religious
wonder
understand
strap
seem
east
article
red
hoe
cancer
suit
cheat
everything
nowhite
thought
wait
obama
away
ppl

INTRODUCTION

- Hate speech is a speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, gender, disability, sexual orientation.

- The law of some countries describes hate speech as speech, gesture or conduct, writing, or display that incites violence or prejudicial action against a protected group or individual on the basis of their membership of the group.

- Social media platforms like Facebook and Twitter has raised concerns about emerging dubious activity such as the intensity of hate, abusive and offensive behavior among us.

Motivation

Potential of social media for spreading hate speech

- 30% internet penetration in India (World Bank, 2016)
- 241 million users of Facebook alone (*The Next Web Report*, 2017)
- 136 million Indians are active social media users (*Yral Report*, 2016)
- 200 million whatsapp users in India (Mashable, 2017)

OBJECTIVE

- The main objective of this work is to develop an automated machine learning based approach for detecting hate speech and offensive language.

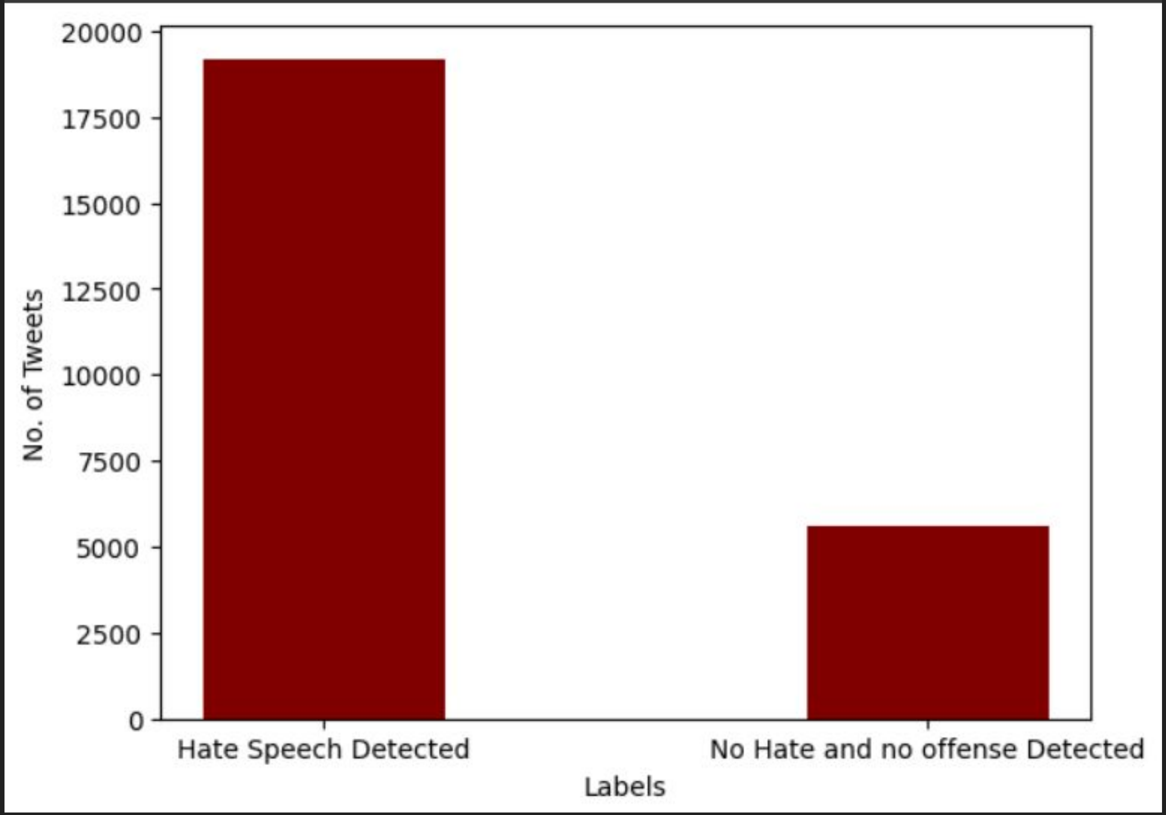
- Automated detection corresponds to automated learning using supervised machine learning technique.

- Classify tweets into Hate speech and No Hate speech no offensive language detection based on number of times tweets contents the word that hurt public sentiments.

The Data:

TWEETS	LABELS
GEEZ..... I think #NorthKorea may be right. #BarackObama is a monkey! Surely acts like one.	1
I, a Catholic and a Jesuit, am grateful for this Muslim holy season for challenging me to be a more dedicated child of God.	1
Saudi can suck a dick tbh, fucking bastards	1
Yes, in barbaric, authoritarian cultures that have since been condemned. Interesting that Christians desire to perpetuate the same idea.	0
Holy shit i am just getting dunked on by a cripple this is unreal.	1
Only cops should have guns!! But... aren't all cops racist and only promote white supremacy and kill blacks on sight? Oh shit, dilemma!!	0

The Data:



- This is a public release of the dataset on kaggle.

- Sourced from Twitter for 24782 combined rows

- The primary outcome variable is the hate speech score
  0 => No hate and no offense speech detected
  1 => hate speech detected

Solution

Create a model that can detect instances of hate speech

Model Types

Decision Tree

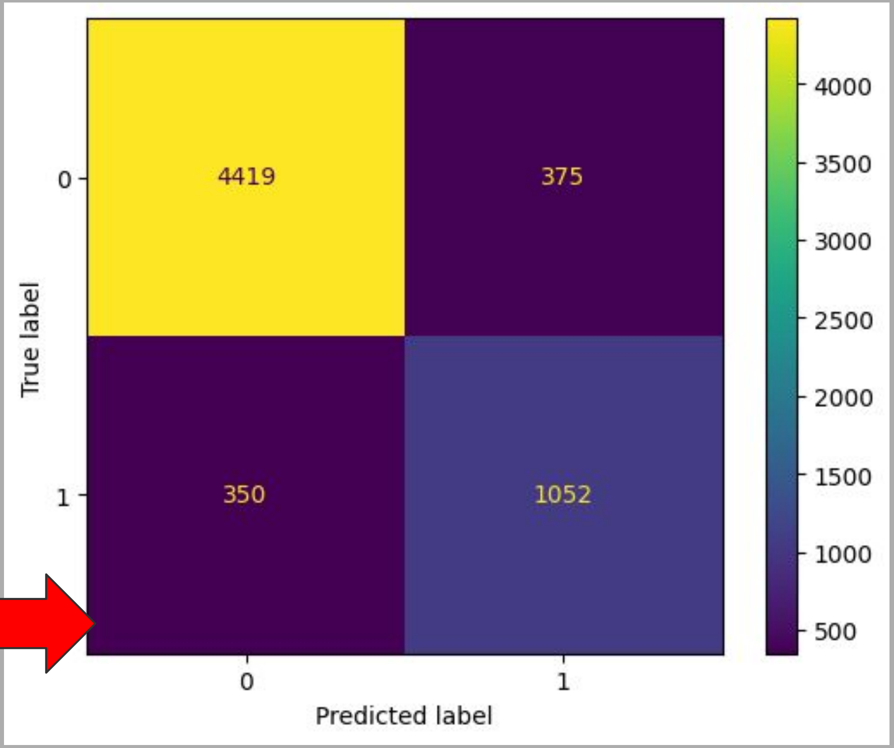
Naive Bayes Classifier

Knn Classifier

Logistic Regression

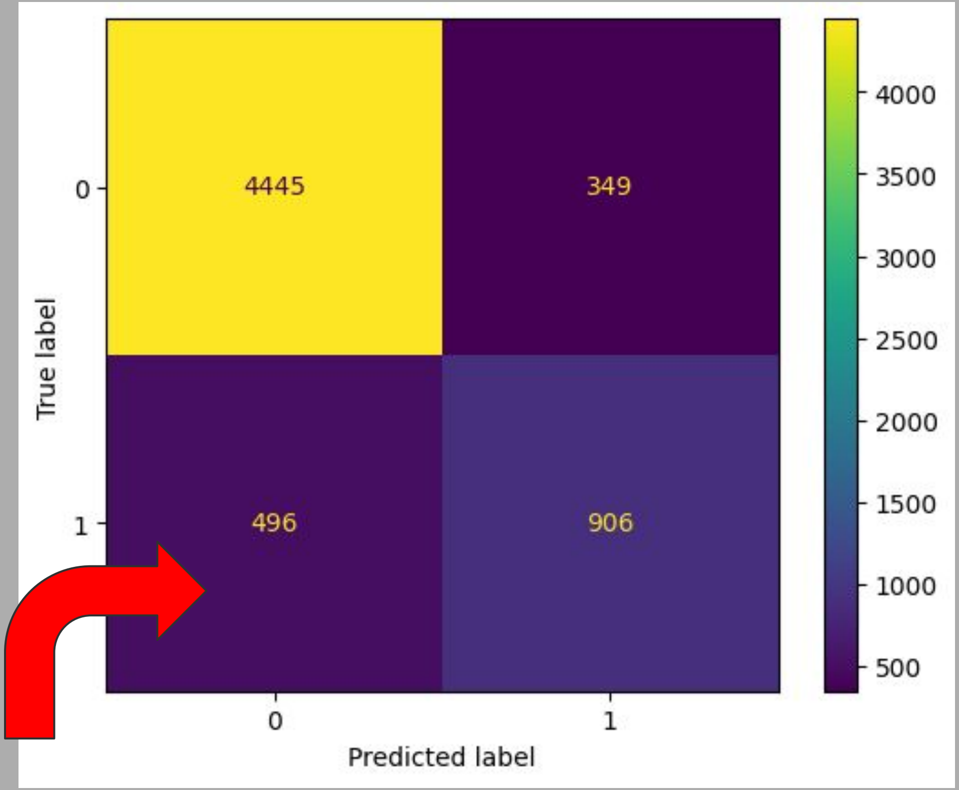
Linear SVM

Implementation DECISION TREE



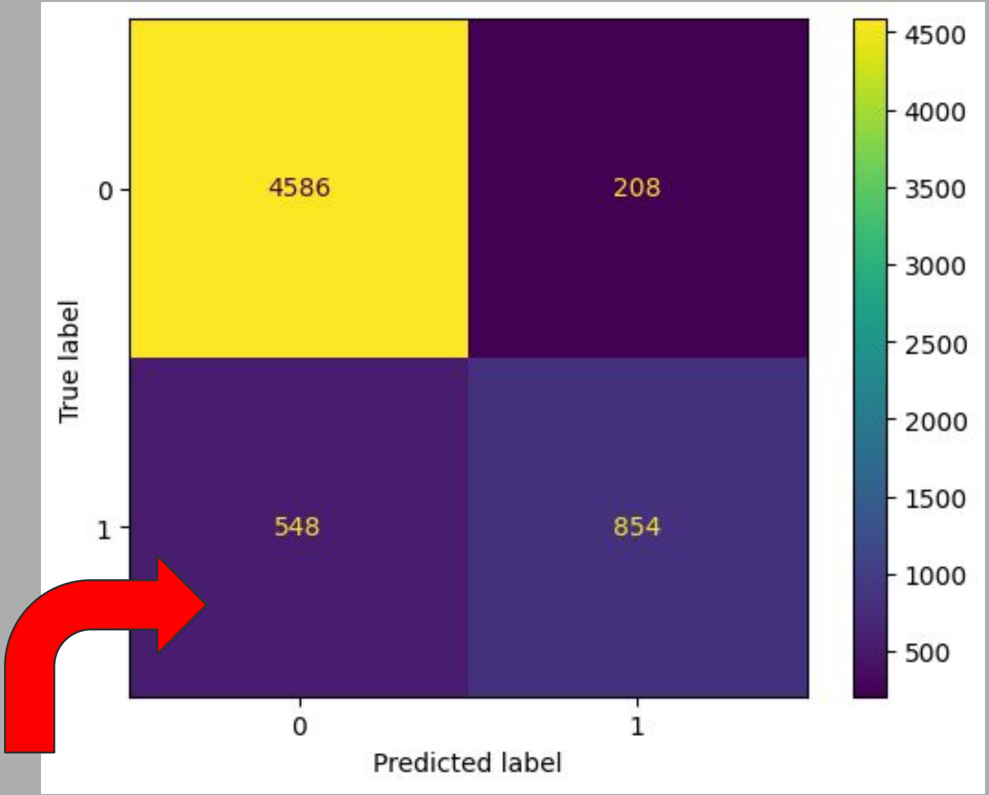
Accuracy	Precision	Recall	F1
0.8826	0.84	0.84	0.84

Implementation K-NEAREST NEIGHBOR



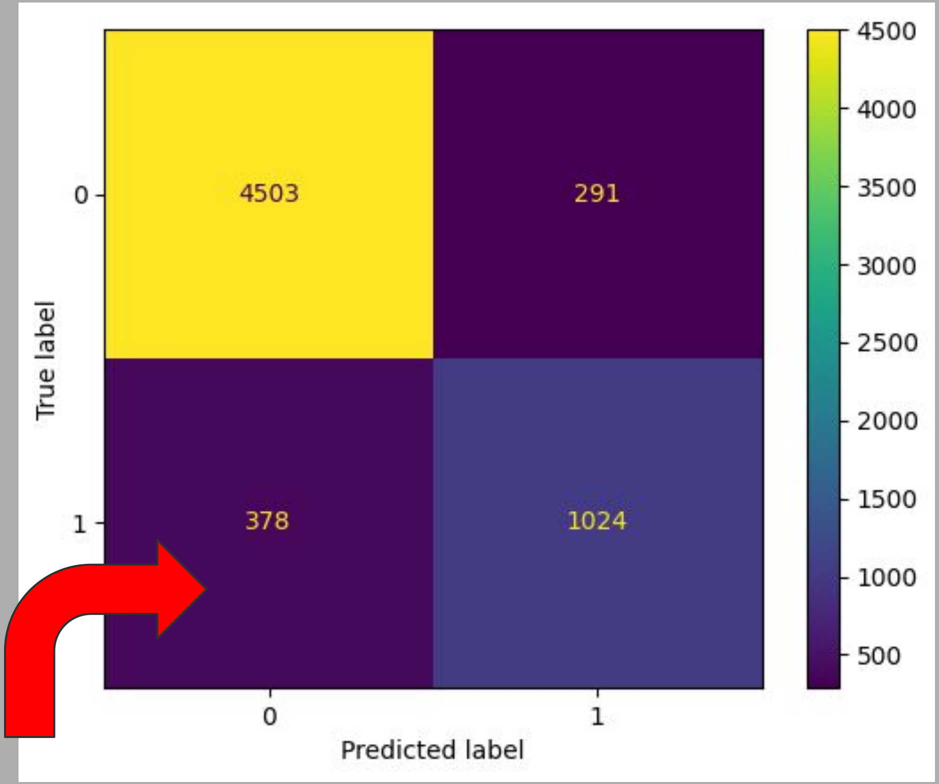
Accuracy	Precision	Recall	F1
0.8636	0.81	0.79	0.795

Implementation NAIVE BAYES



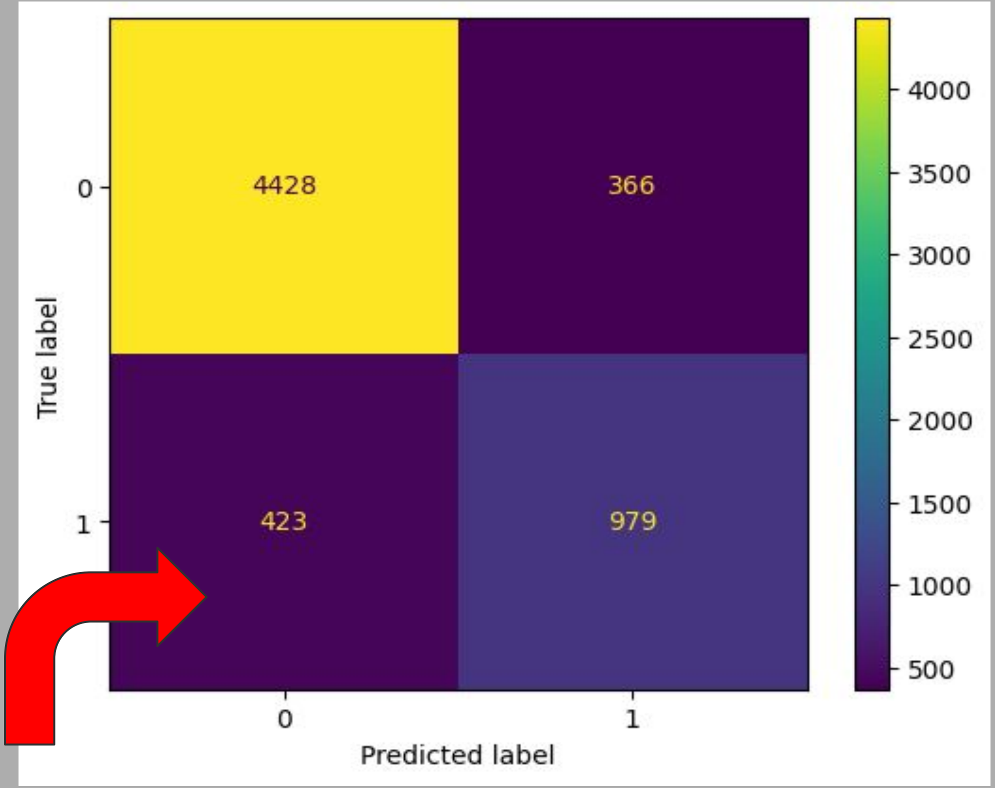
Accuracy	Precision	Recall	F1
0.8779	0.8450	0.7850	0.8050

Implementation LOGISTIC REGRESSION



Accuracy	Precision	Recall	F1
0.8920	0.85	0.8350	0.84

Implementation LINEAR SVM



Accuracy	Precision	Recall	F1
0.8726	0.85	0.8350	0.84

The Predictions:

Tweets (UNSEEN DATA)	Labels	Prediction
My God look at these BARBARIC PPL. And these pigs all watching. They r a very, very sick culture. And Omar and Rashid want this crap here. This is America not the Middle East. Round them up and send them home.	1	1
That's right would have to be at least two more because someone is going to need to shut your whore mouth up while my cock rapes your asshole over and over again while daddy's other friend is in your slutty little cunt.	1	1
Instead of killing americans just kill brown kids instead	1	0
I love that you enjoy being trans.	0	0
Fucking panty waste UK politicians can suck a dick! They sold out their own people for a shit god, preferential rights for moslems. Look at your own policies, calling Trump anything just shows your own dumbmotherfuckery. Stand up UK citizens and take your fucking country back! For nearly 2000 years you've fought off every invader in the world but now your politicians have gone ass up for allah. Some of the greatest kings to ever walk the earth you've decended from, they'd roll over in their graves if they knew	1	1

FUTURE WORK:

- Increasing record size for greater accuracy.
- Create a Chrome extension that can detect hate speech and censor it.
- Create a social media bot i.e instagram/twitter/facebook that can detect and flag hate speech.
- Use text summary generation to replace hate speech text with a censored summary.

THANK YOU