Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Aggressive Language Classifier

Natural Language Processing

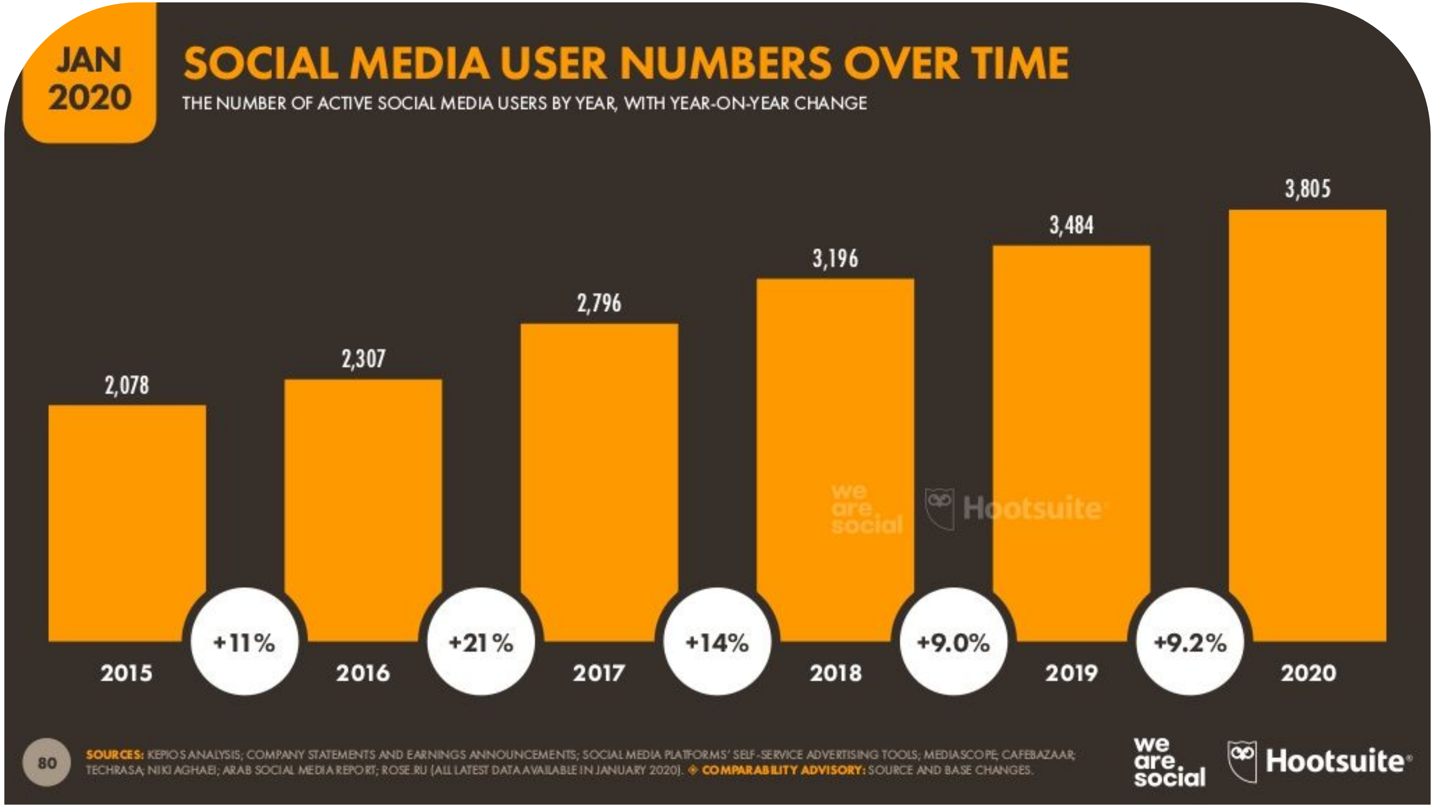
Date: July 1st, 2020

Team: Manoj Kolpe Lingappa and Alan Gomez

Github: https://github.com/alanorlando95/Aggressive-language-classifier

Google colab： https://colab.research.google.com/drive/1AMwk1apDr-WZaKd66H00MfFB2XJcwUAU?usp=sharing

Motivation



SOURCES: KEPIOS ANALYSIS; COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; MEDIASCOPE; CAFEBAZAAR; TECHRASA; NIKI AGHAEI; ARAB SOCIAL MEDIA REPORT; ROSE.RU (ALL LATEST DATA AVAILABLE IN JANUARY 2020). ◈ COMPARABILITY ADVISORY: SOURCE AND BASE CHANGES.

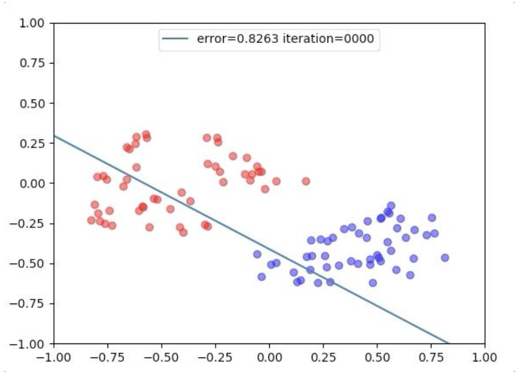
Fig 1. Social media users over time (in millions) [1].

Motivation

- 87 percent of young people in the USA have seen cyberbullying occurring online (2017 - 2019) [2].
- 36.5 percent of people in the USA feel they have been cyberbullied in their lifetime (2017 - 2019) [2].

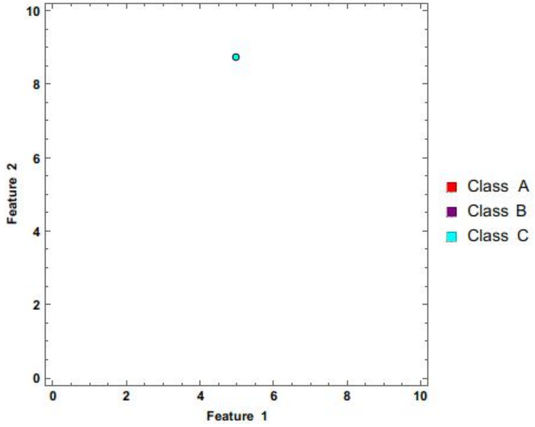
Modern problems require modern solutions (maybe not too modern)

- Logistic Regression
- Naive Bayes
- Support Vector Machine



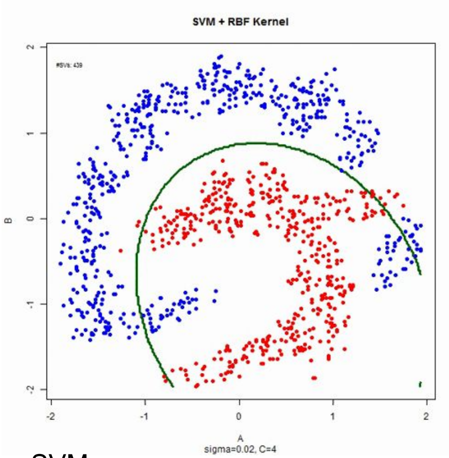
Logistic regression

Credits: https://brainbomb.org/Artificial-Intelligence/Machine-Learning/ML-Linear-Classification-Logistic-Regression/



Naive bayes

Credits: https://commons.wikimedia.org/wiki/File:Naive_Bayes_Classifier.gif

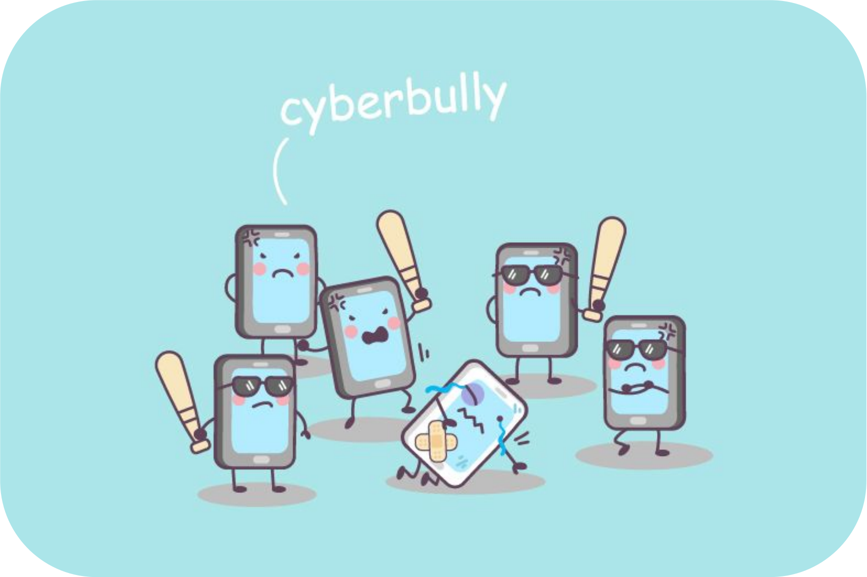


SVM

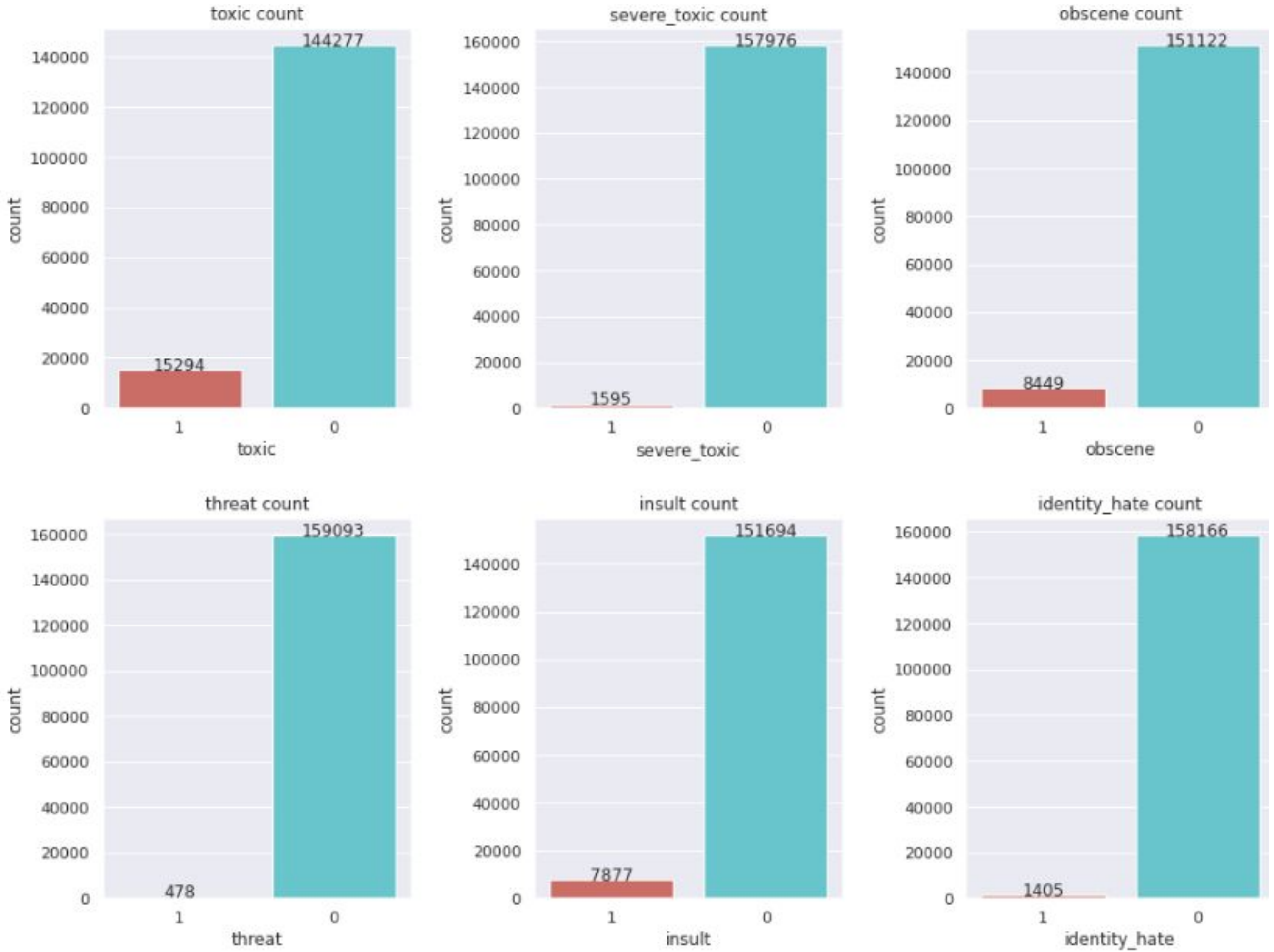
Credits: https://gfycat.com/sentimentalthickbullfrog

Dataset

- Jigsaw Multilingual Toxic Comment Classification, made up of English comments from Wikipedia’s talk page edits [3].



Credits: https://mc.ai/detecting-toxic-comment/

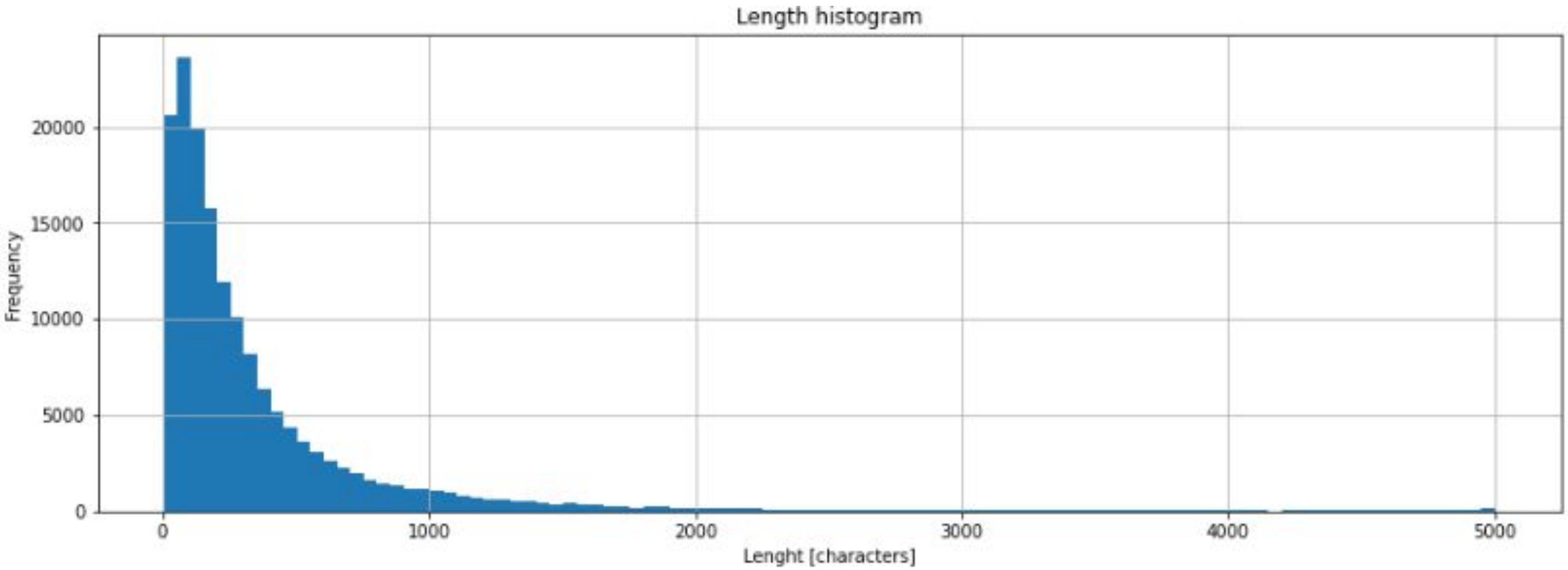

toxic count
count
140000
120000
100000
80000
60000
40000
20000
0
15294
144277
1
0
toxic
severe_toxic count
count
160000
140000
120000
100000
80000
60000
40000
20000
0
1595
157976
1
0
severe_toxic
obscene count
count
140000
120000
100000
80000
60000
40000
20000
0
8449
151122
1
0
obscene
threat count
count
160000
140000
120000
100000
80000
60000
40000
20000
0
478
159093
1
0
threat
insult count
count
140000
120000
100000
80000
60000
40000
20000
0
7877
151694
1
0
insult
identity_hate count
count
160000
140000
120000
100000
80000
60000
40000
20000
0
1405
158166
1
0
identity_hate

Toxic words

fuck
SUCK
Wikipedia
shit
NIGGER
fucking
faggot
page
will
people
ass
JEW
dick
CUNT
bitch
penis
cock
one
hate
DIE
stupid
FAT
know
moron
article
Go
gay
block
user
idiot
asshole
time
life
love
admin
Bullshit
stop
take
make
now
back
bad
even
want
talk
think
going
piece
wiki
eat
Bark
really
hi
vandalism
WANKER
BALL
FAG
PIG
fucksex
LOSER
FUCKER
hell
nothing
Nipple
AID
COCKSUCKER
ANAL
KILL
DAMN
SHUT
DONT
FREEDOM
care
blocked
little
guy
said
know
say
give
come
mean
delete
comment
source
dickhead
mother
nigga
Bastered
BUTTSECKS
Hey
look
keep
MOTHJER
DOG
information
first
editor
tell
fact
poop
change
administrator
dumb
SUPER
year
heil
hitler
nazi
homo
rule
nameediting
MEXICAN
BUSH
Please
something
reason
old
well
way
need
never
banned
site
world
post
better
real
work
big
hope
many
leave
word
BOOB
HUGE
NOOB
time
new
hell
use
try
warning
Thank
still
let
much
fool
live
trying
personal
anything
someone
bitches
racist
must
Pro
thing
got
read
wrong
made
crap
put
IP
GOD

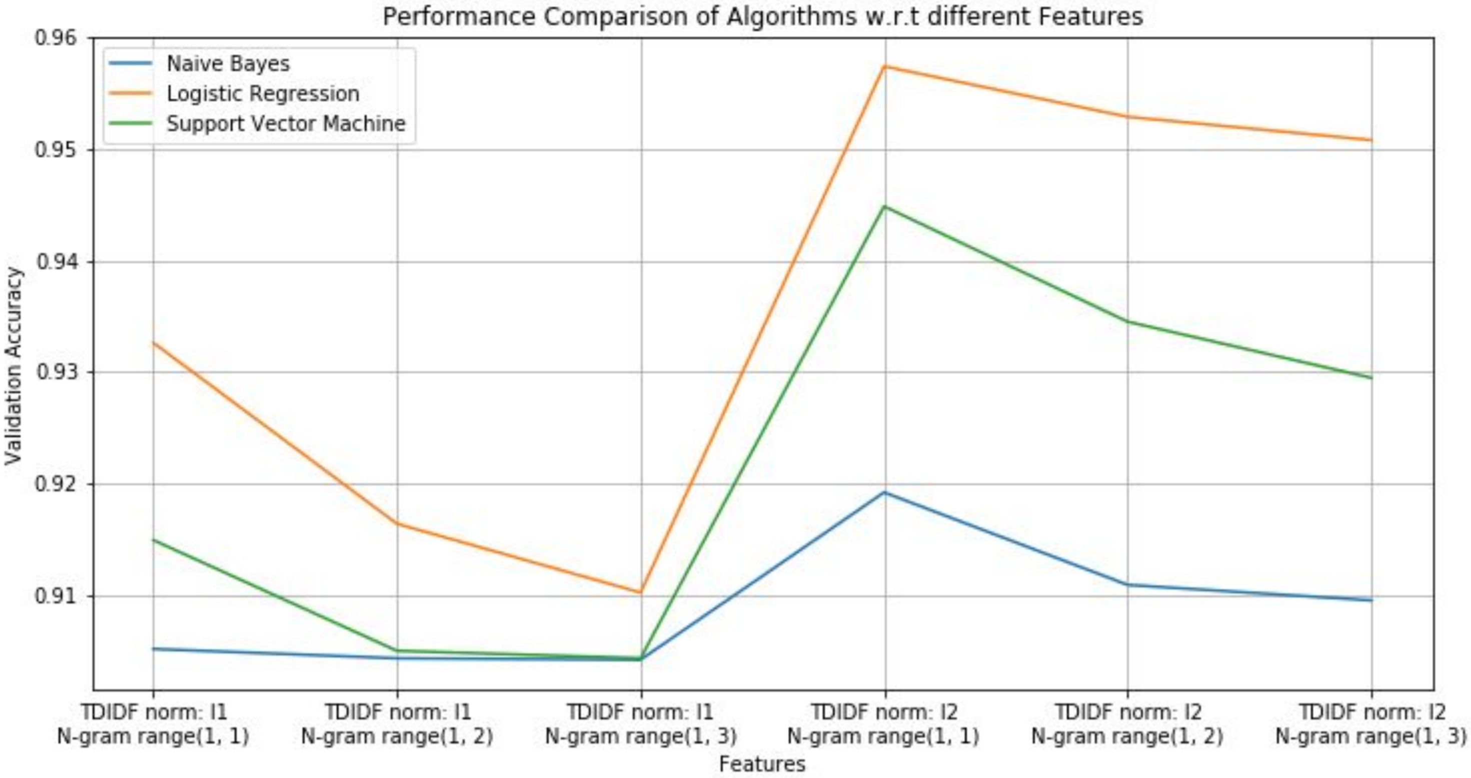
Non toxic words

article
Wikipedia
page
Please
talk
edit
Thank
source
think
one
will
know
time
see
use
user
information
deletion
name
link
look
subject
two
year
list
find
editor
say
good
right
take
may
first
state
removed
content
change
well
seem
issue
help
question
people
make
work
discussion
read
even
fact
image
reference
reason
need
way
comment
new
blocked
thing
much
want
point
word
problem
someone
now
still
note
back
case
block
without
wiki
using
history
go
man
believe
hope
sorry
leave
better
fair
statement
best
account
come
agree
sure
really
group
day
copyright
keep
put
tag
language
must
claim
number
original
other
actually
far
ask
rather
continue
stop
post
anyone
something
added
never
deleted
guideline
WP
making
template
might
made
opinion
Hi
let
style
vandalism
person
found
contribution
feel
place
add
welcome
part
text
lot
speedy
English
policy
message
already
editing
thought
support
wrong
realize
review
material
request
though
said
notable
world
Hello
little
rule
try
created
adding
nothing
going
matter
non
top
done
reliable
view
term
least
revert
title
website
remove
understand
mean
long
site
simply
example
personal
give
last
different
consensus
another
probably
trying
free
section
UTC
either
great
admin
u
anything
write


Length histogram
Frequency
0
5000
10000
15000
20000
0
1000
2000
3000
4000
5000
Lenght [characters]

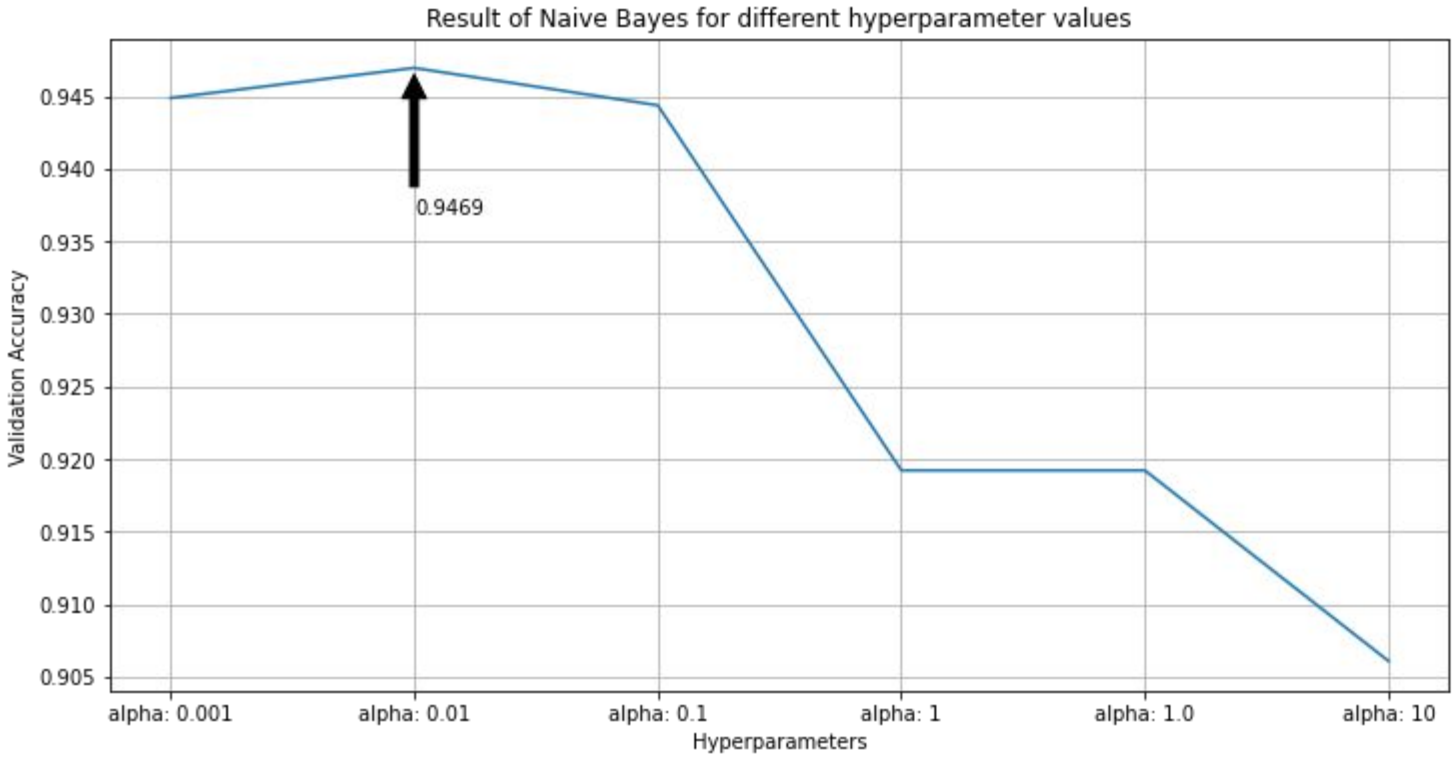
The implementation based on Gaydhani et. al [4]

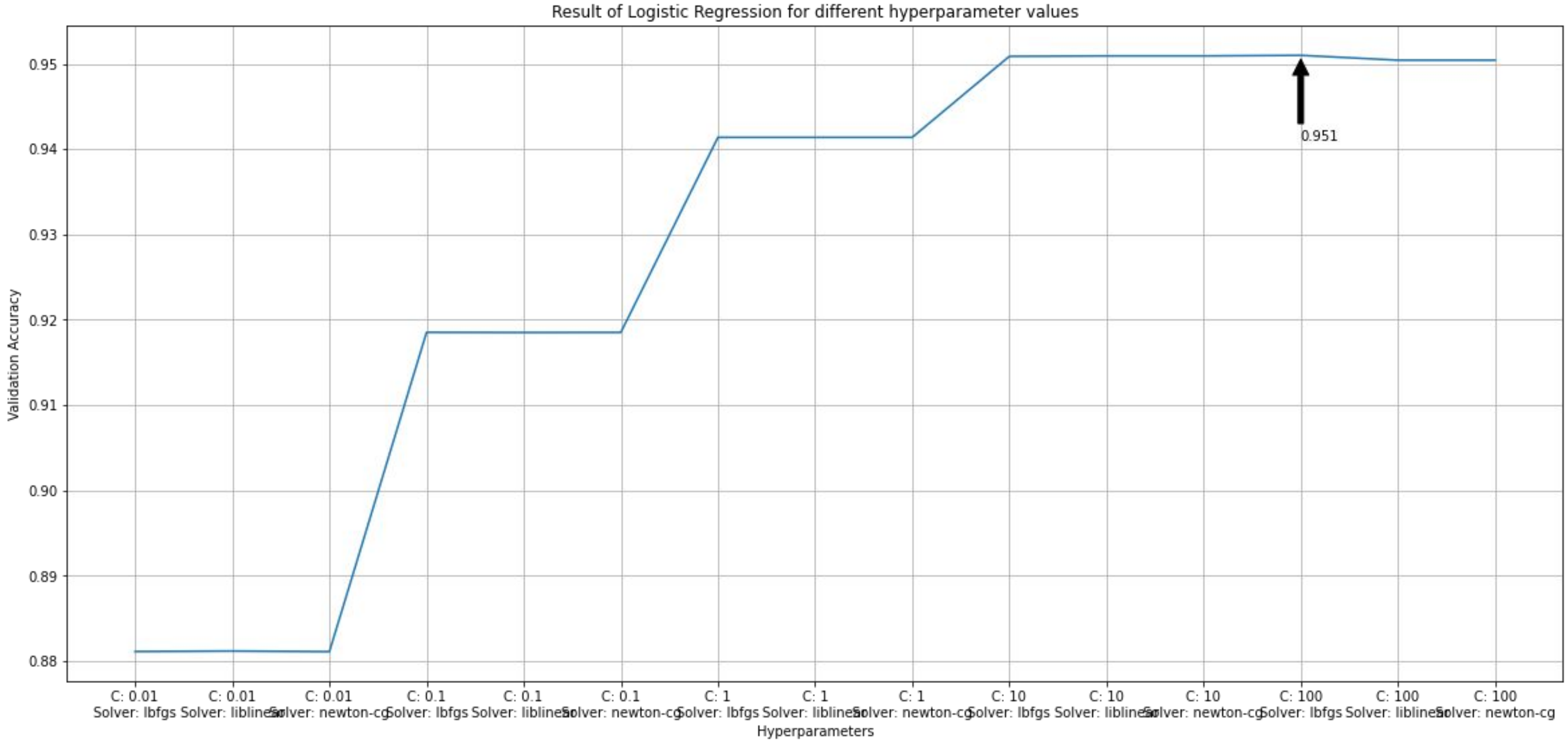
- Train models by performing greedy search with different parameters:
  - CountVectorizer:
    - n-gram range ((1, 1), (1, 2), and (1, 3))
  - TfidfTransformer:
    - norm (L1 and L2)
- CountVectorizer extracts the n-gram features
- TfidfTransformer weight the n-gram features

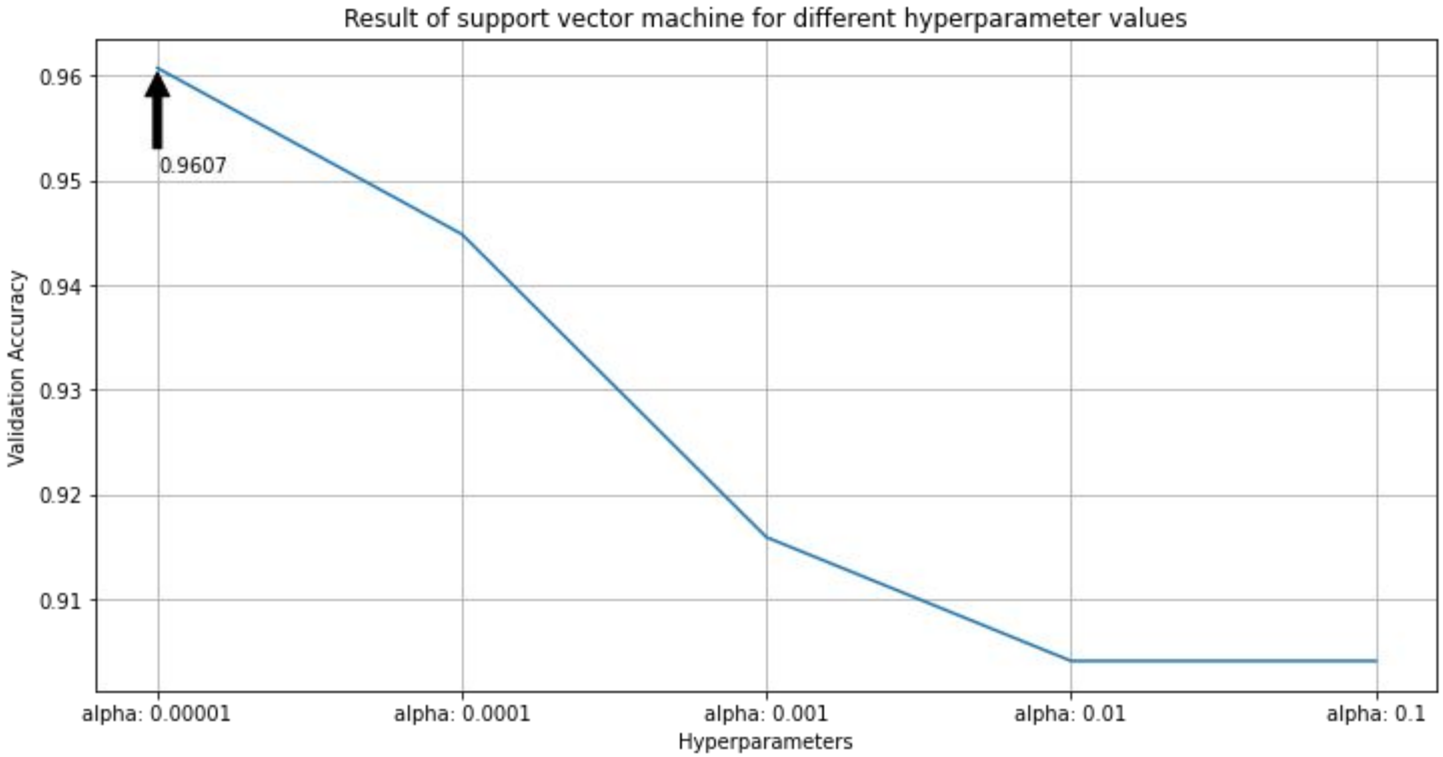

Performance Comparison of Algorithms w.r.t different Features
Naive Bayes
Logistic Regression
Support Vector Machine
Validation Accuracy
0.96
0.95
0.94
0.93
0.92
0.91
TDIDF norm: l1
N-gram range(1, 1)
TDIDF norm: l1
N-gram range(1, 2)
TDIDF norm: l1
N-gram range(1, 3)
TDIDF norm: l2
N-gram range(1, 1)
TDIDF norm: l2
N-gram range(1, 2)
TDIDF norm: l2
N-gram range(1, 3)
Features

The implementation based on Gaydhani et. al [4]

- Compare models
- Hyperparameter tuning:
    - MultinomialNB:
        - Additive (Laplace/Lidstone) smoothing parameter (0.01, 0.1, 1, and 10)
    - LogisticRegression:
        - Inverse of regularization strength (10 and 100)
        - Solver (newton-cg, liblinear, and saga)
    - SVM:
        - Scale of regularization term (0.00001, 0.0001, 0.001, 0.01, and 0.1)


Result of Naive Bayes for different hyperparameter values
Validation Accuracy
0.945
0.940
0.935
0.930
0.925
0.920
0.915
0.910
0.905
0.9469
alpha: 0.001
alpha: 0.01
alpha: 0.1
alpha: 1
alpha: 1.0
alpha: 10
Hyperparameters

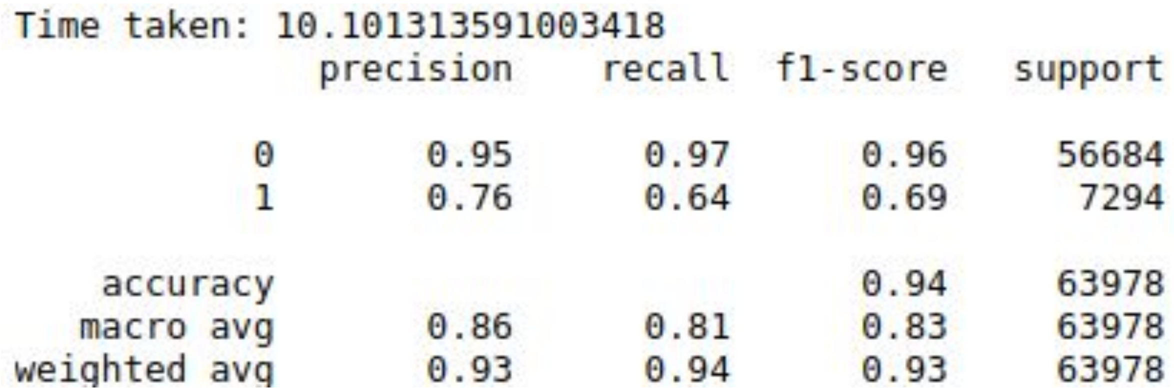

Result of Logistic Regression for different hyperparameter values
Validation Accuracy
0.95
0.94
0.93
0.92
0.91
0.90
0.89
0.88
0.951
C: 0.01 Solver: lbfgs
C: 0.01 Solver: liblinear
C: 0.01 Solver: newton-cg
C: 0.1 Solver: lbfgs
C: 0.1 Solver: liblinear
C: 0.1 Solver: newton-cg
C: 1 Solver: lbfgs
C: 1 Solver: liblinear
C: 1 Solver: newton-cg
C: 10 Solver: lbfgs
C: 10 Solver: liblinear
C: 10 Solver: newton-cg
C: 100 Solver: lbfgs
C: 100 Solver: liblinear
C: 100 Solver: newton-cg
Hyperparameters


Result of support vector machine for different hyperparameter values
Validation Accuracy
0.96
0.95
0.94
0.93
0.92
0.91
0.9607
alpha: 0.00001
alpha: 0.0001
alpha: 0.001
alpha: 0.01
alpha: 0.1
Hyperparameters

Total data points: 156871
Type of comment: Toxic

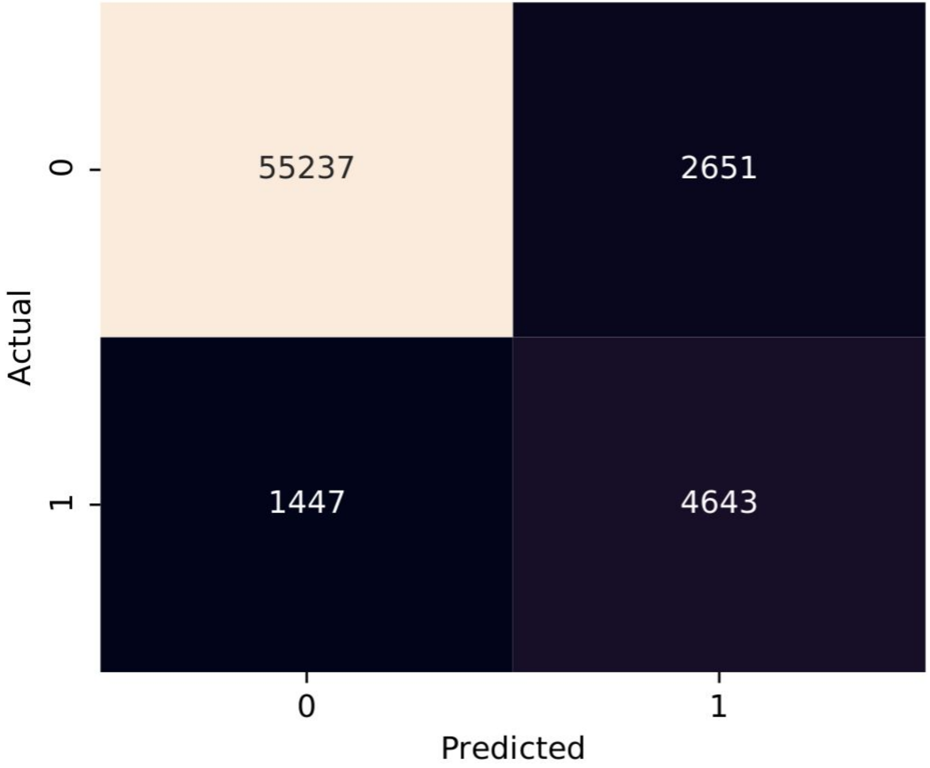
Google colab RAM: 12GB	Time taken to train w.r.t different features (min)	Time taken for Hyperparameter Tuning (min)
Naive Bayes	42	7
Logistic regression	81	35
Support vector machine	52	10

SVM results (toxic)

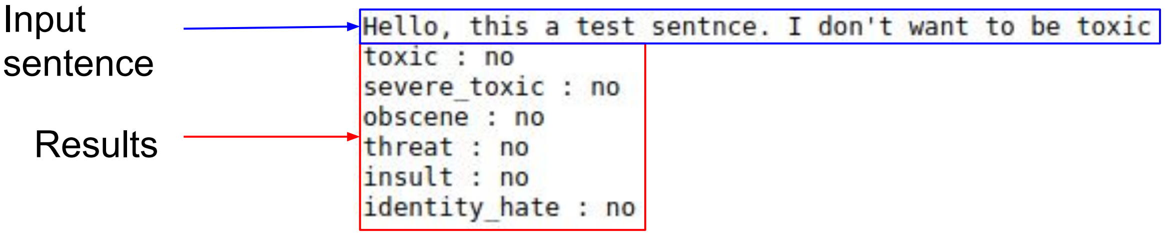


```
Time taken: 10.101313591003418
              precision    recall  f1-score   support

           0       0.95      0.97      0.96     56684
           1       0.76      0.64      0.69      7294

    accuracy                           0.94     63978
   macro avg       0.86      0.81      0.83     63978
weighted avg       0.93      0.94      0.93     63978
```

Confusion matrix (toxic)



Train the best model to classify the data into the other categories of the dataset

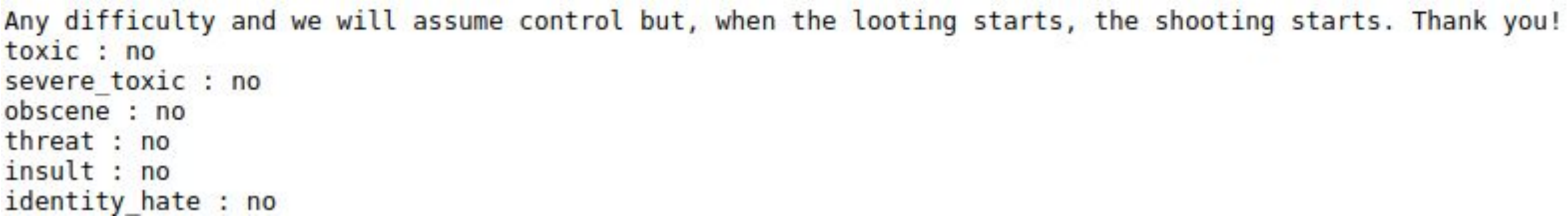


Input sentence

Results

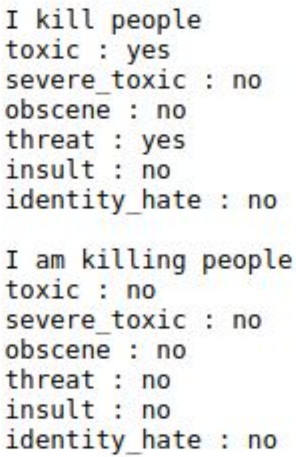
```
Hello, this a test sentnce. I don't want to be toxic
toxic : no
severe_toxic : no
obscene : no
threat : no
insult : no
identity_hate : no
```

Implicit threats are not detected



```
Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you!
toxic : no
severe_toxic : no
obscene : no
threat : no
insult : no
identity_hate : no
```

Sensitive to unknown words



```
I kill people
toxic : yes
severe_toxic : no
obscene : no
threat : yes
insult : no
identity_hate : no

I am killing people
toxic : no
severe_toxic : no
obscene : no
threat : no
insult : no
identity_hate : no
```

Take Home Message

- The training data can influence the results on a classification task
  - Number of positive and negative examples
  - Unseen words
- One can train different models comparing different parameters using sklearn
- Different TFIDF norm and n gram range affects the validation accuracy.
- Hyperparameter tuning can give better validation accuracy than using gridsearch with respect to different features.
- Confusion matrix can be used to describe performance of a classification model.
- Traditional solver such as naive bayes, SVM and logistic regression produce reasonable performance in classification of toxic comment.

References

[1] Kemp, Simon. “Digital 2020: Global Digital Overview - DataReportal – Global Digital Insights.” DataReportal. DataReportal – Global Digital Insights, January 30, 2020. https://datareportal.com/reports/digital-2020-global-digital-overview.

[2] “51 Critical Cyberbullying Statistics in 2020.” BroadbandSearch.net. Accessed June 18, 2020. https://www.broadbandsearch.net/blog/cyber-bullying-statistics.

[3] “Jigsaw Multilingual Toxic Comment Classification.” Kaggle. Accessed June 18, 2020. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data.

[4] Gaydhani, Aditya, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv:1809.08651 (2018).