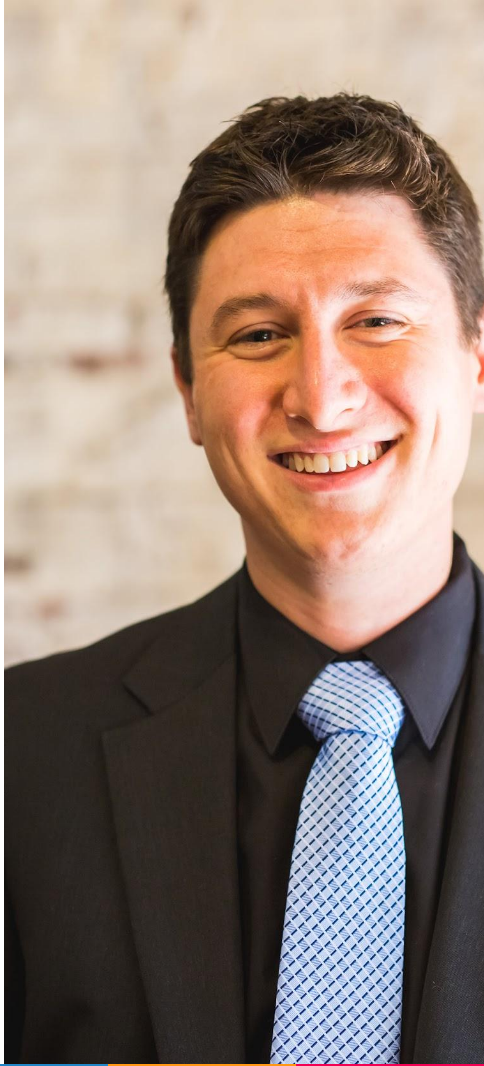
reddit Classification

Hello!

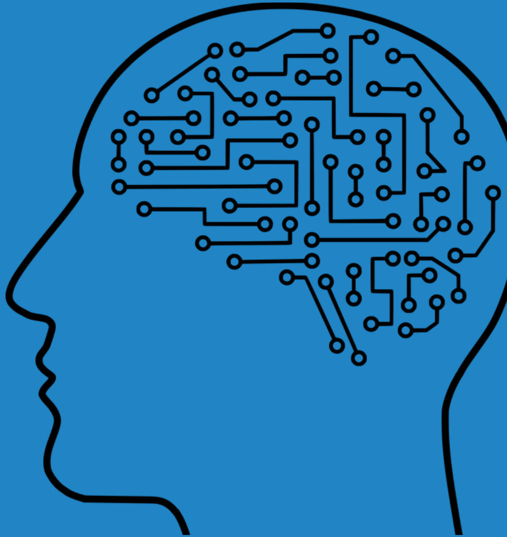
I am Alex Fioto

I am here to talk about classifying subreddit posts using scikit-learn classifier models



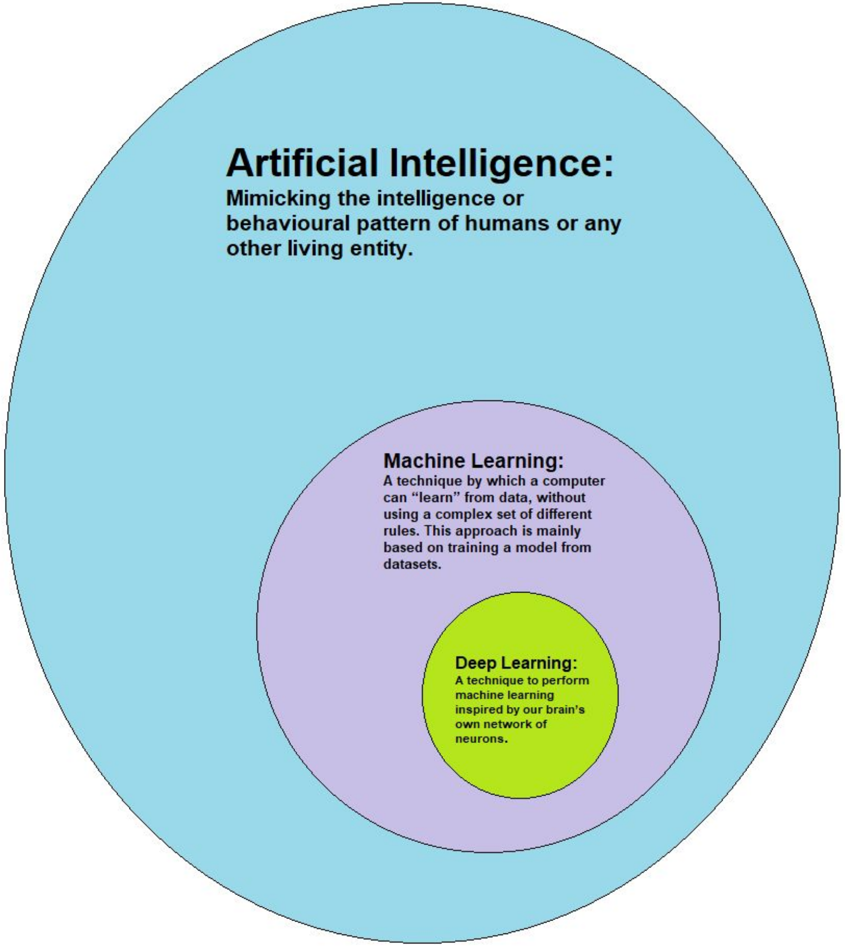
Problem:

Trained on data from two similar subreddits, how well can a model classify a Reddit post?



Machine Learning vs Artificial Intelligence

How similar?



Artificial Intelligence:
Mimicking the intelligence or behavioural pattern of humans or any other living entity.

Machine Learning:
A technique by which a computer can “learn” from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.

Deep Learning:
A technique to perform machine learning inspired by our brain’s own network of neurons.

Data Collection

	Subreddit	Rows
Machine Learning	**r/MachineLearning/**	**9743**
Artificial Intelligence	**r/ArtificialInteligence/**	**9998**

Text Preprocessing

Feature Creation

Combining all text from each post to create an "all_text" feature

Lemmatizing

Create custom lemmating function and apply to "all_text"

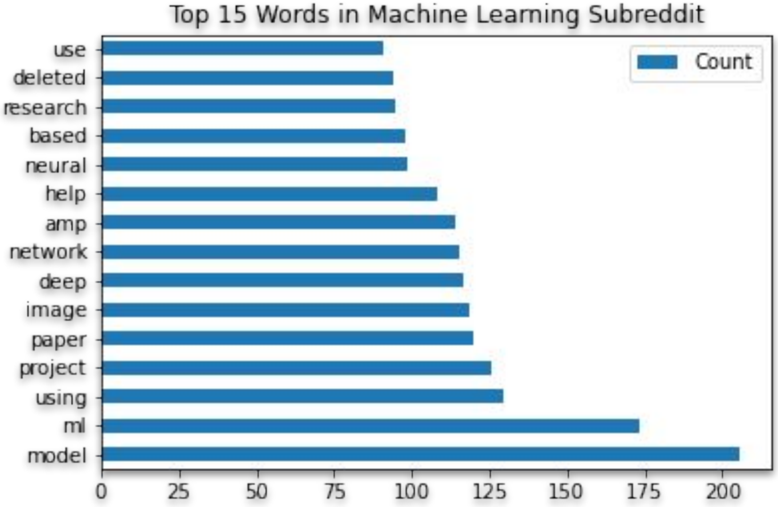
Stemming

Create custom stemming function and apply to "all_text"

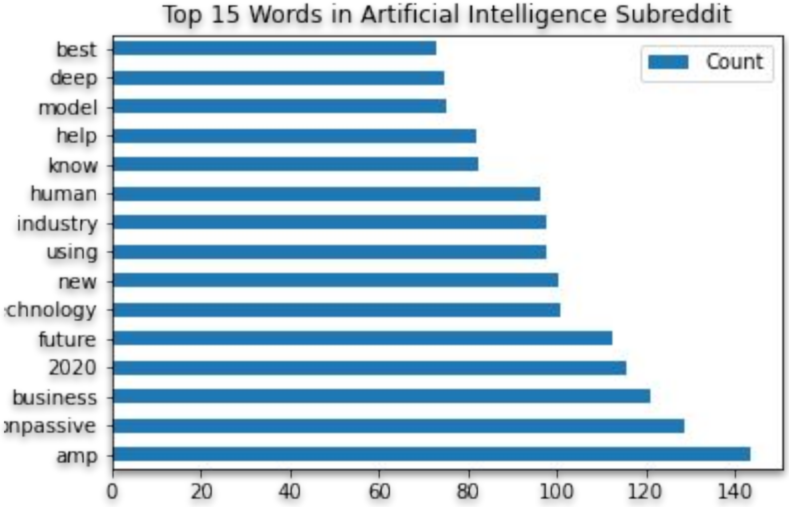
machine
model
data
artificial
using
deep
time
paper
new
network
image
know
human
work
use
one
help
neural
people
code
make
learn
way
feature
result
good
year
series
library
give
well
tool
robot
trying
gpt-3
start
around
trained
doe
need
first
even
user
test
topic
world
week
see
real
set
job
now
may
etc
let
look
right
made
lot
every
layer
face
idea
two
get
post
field
system
dataset
business
google
really
different
course
anyone
better
create
method
working
output
point
input
task
take
find
free
vision
book
solution
thought
text
state
found
train
top
number
next
app
feel
large
simple
much
x200b
just
best
think
based
future
thing
build
game
example
researcher
source
similar
covid-19
pytorch
read
want
also
video
used
many
word
thank
science
application
github
predict
come
looking
please
analytics
since
small
value
website
seems
domain
datasets
change
program
current
reading
sample
market
deleted
improve
another
last
part
online
team
nlp
link
case
show
talk
mean
say
class
space
bot
still
list
high
role
specific
gpu
day
reddit
news
share
speech
environment
resource
issue
expert
design
long
across
insight
detail
quick
built
single
natural
voice
order
tried
check
level
going
useful
end
loss
error
kind
created
author
person
making
mind
cost
put
recent
little
already
hard
answer
version
pattern
today
digital
tutorial
random
cloud
actually
mobile
maybe
audio
label
math
reason
focus
might
trend
blog
quality
content
study
someone
basic
advanced
place
wholly
master
second
pretty
concept
done
search
impact
getting
strategy
instead
theory
graph
great
given
social
type
able
arxiv
sure
guy
bias
memory
career
love
due
bit
file
size
try
run
life
live
full
big
open
term
power
tech
step
funny
related
generation
difference
keep
write
area
major
india
goal
weight
group
advance
several
medium
sequence
specifically

I like data!

Common words



Top 15 Words in Machine Learning Subreddit



Top 15 Words in Artificial Intelligence Subreddit

ONPASSIVE

www.onpassive.com

https://readwrite.com/2019/02/11/a-comprehensive-guide-to-google-accelerated-mobile-pages/

Classifiers Used:

Naive-Bayes

- Simple
- Quick to train
- Interpretable

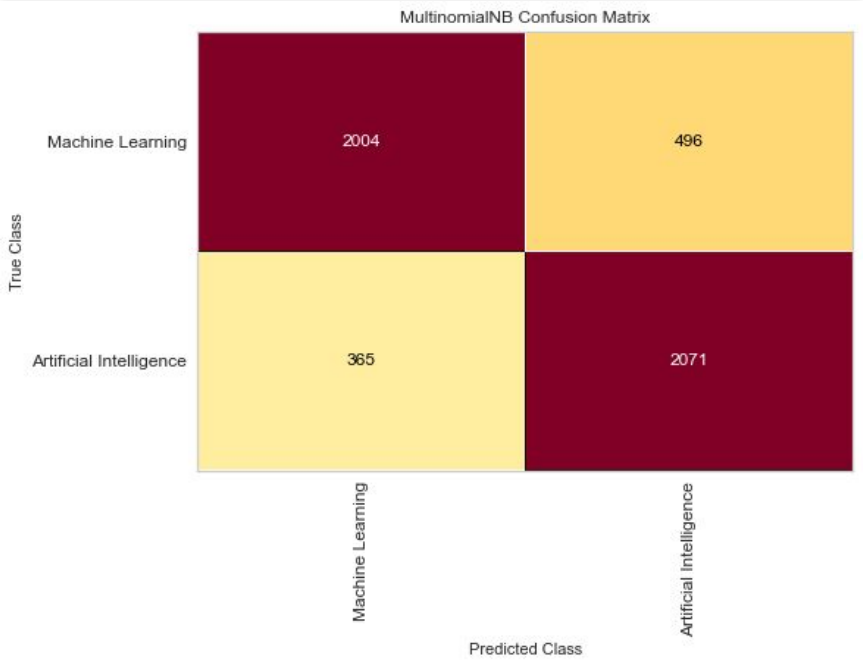
Random Forest

- Powerful
- Interpretable
- Accurate

Classification Metrics

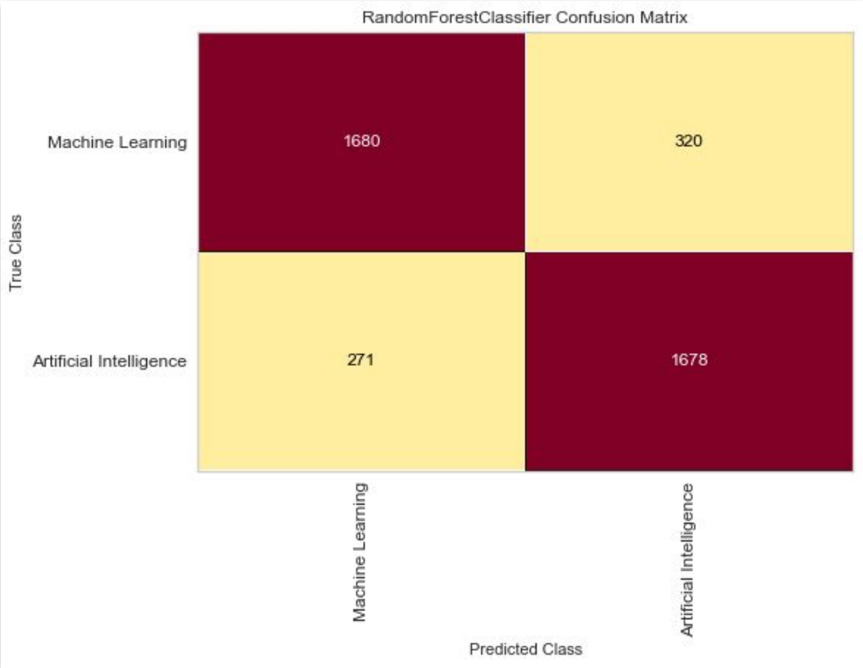
- Accuracy
- F-1 Score
- Confusion Matrix

Confusion Matrices



MultinomialNB Confusion Matrix

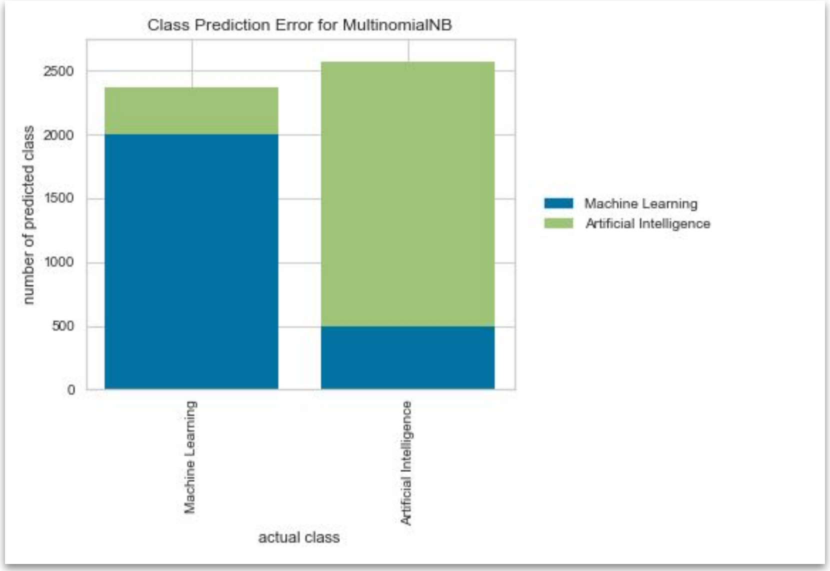
	Machine Learning	Artificial Intelligence
Machine Learning	2004	496
Artificial Intelligence	365	2071

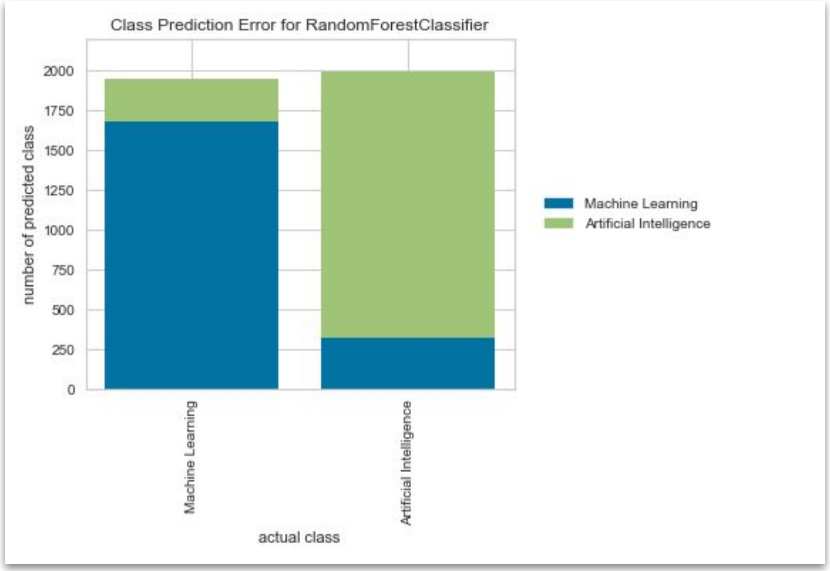


RandomForestClassifier Confusion Matrix

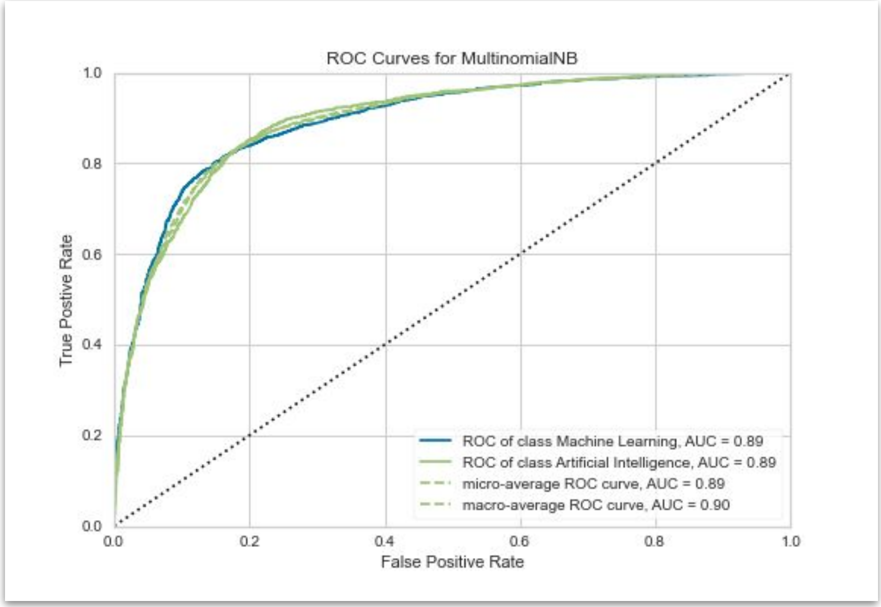
	Machine Learning	Artificial Intelligence
Machine Learning	1680	320
Artificial Intelligence	271	1678

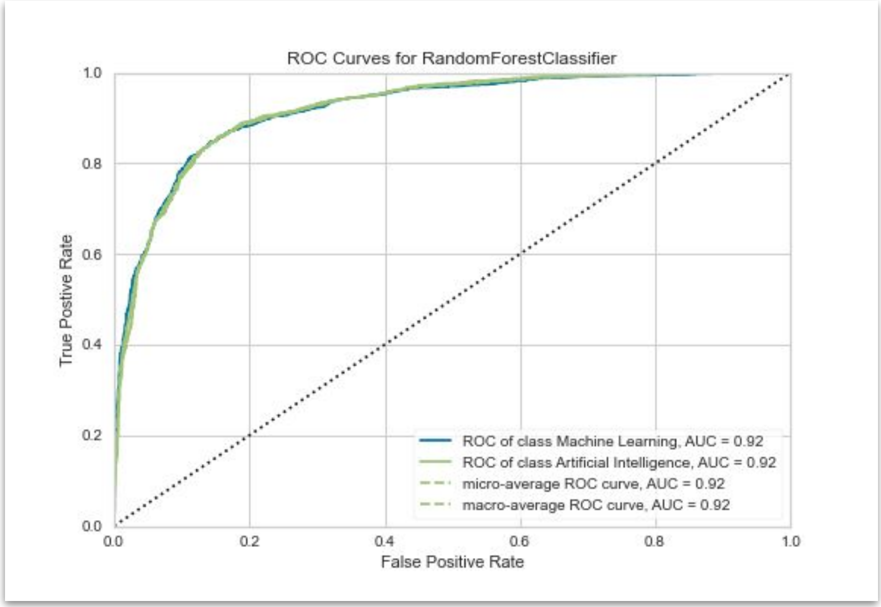
Class Prediction Error





ROC Curves


ROC Curves for MultinomialNB
True Postive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class Machine Learning, AUC = 0.89
ROC of class Artificial Intelligence, AUC = 0.89
micro-average ROC curve, AUC = 0.89
macro-average ROC curve, AUC = 0.90


ROC Curves for RandomForestClassifier
True Postive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class Machine Learning, AUC = 0.92
ROC of class Artificial Intelligence, AUC = 0.92
micro-average ROC curve, AUC = 0.92
macro-average ROC curve, AUC = 0.92

Model Results

	Test Accuracy	Train Accuracy	Test F-1 Score	Train F-1 Score
Naive-Bayes	**0.8256**	**0.8321**	**0.8279**	**0.8353**
Random Forest	**0.8519**	**0.9851**	**0.8513**	**0.9850**

Hyperparameters Used:

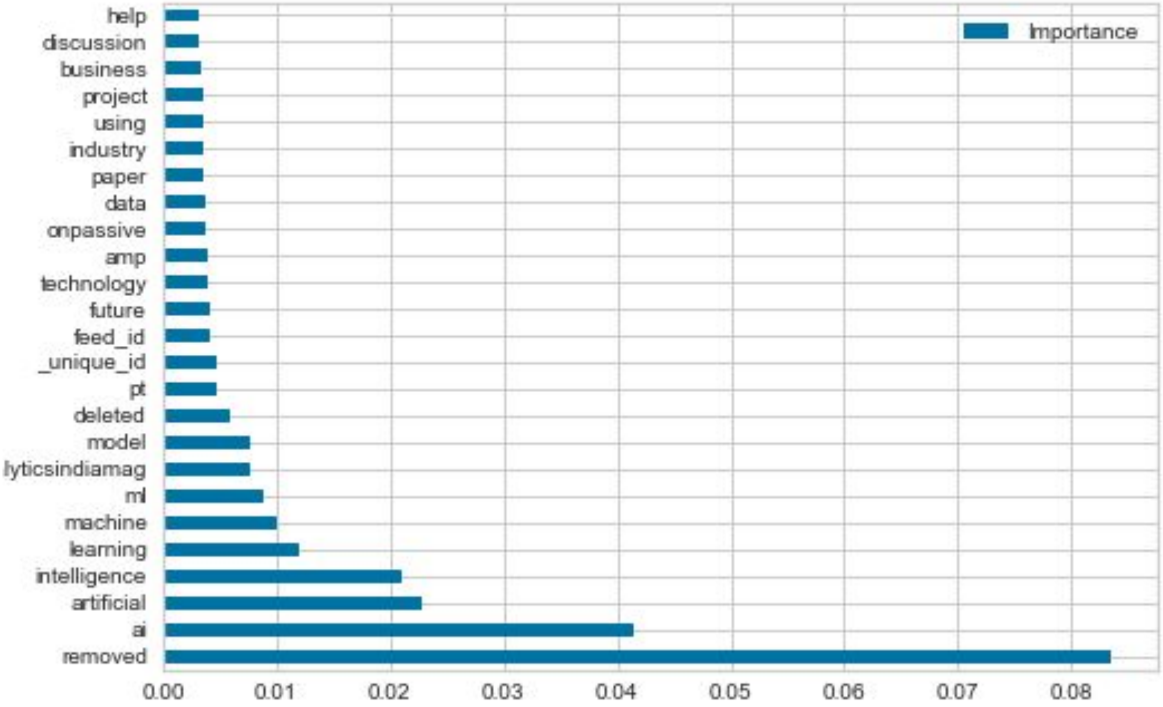
Naive-Bayes

- Tfidfvectorizer
- alpha=250
- fit_prior=True

Random Forest

- Tfidfvectorizer
- n_estimators=500
- criterion='entropy'
- max_depth=None
- min_samples_split=2
- min_samples_leaf=1
- max_features='auto'
- max_leaf_nodes=None
- min_impurity_decrease=0.0

Random Forest Feature Importance



Conclusion

Summary

- Random Forest
- Many features
- 85% Accuracy

Limitations

- Narrow view
- “Removed” Dependent

Future

- Collect more data
- Logistic Regression
- Neural Net

Thanks!

Any questions?