NLP & Reddit API

Manuel Molina

Project 3

Overview

1. Background Information
2. Problem Statement
3. Data
4. Model Methods
5. Analysis
6. Conclusion & Recommendations

Reddit

- "Network of communities where people can dive into their interests, hobbies and passions"
- Collection of forums where people can share content

No Stupid Questions (NSQ)

"for people to feel free to ask the questions they might be embarrassed or ashamed to ask elsewhere"

1.5m members; established in Aug 2011

"Where do STI's originate from?

Too Afraid To Ask (TATA)

"everything and anything you were too afraid to ask"

2.6m members; established Feb 2013

"How sudden is Sudden Infant Death Syndrome (SIDS)?"

Background

Problem

Given a post, can I identify with above .80 accuracy which subreddit the post belongs to using NLP?

Data

Data Collection

Web Scraped Reddit API
- title, selftext, created_utc

199,737 Total Posts (reduced to 122,978)
- 99,849 NSQ Posts (reduced to 60,733)
- 99,888 TATA Posts (reduced to 62,245)

Data Used

Exclusions:
- [removed], [deleted], and posts without self-text

Tested on random sample of 40,000 posts
Baseline accuracy: .5028 (TATA) & .4971 (NSQ)

Model Methods

Model Tests

- 9 Models
- 1 Random Search
- 1 Bayesian Search

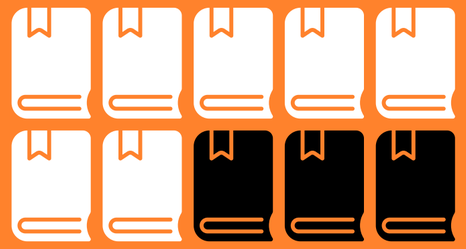
Model	Train	Test
Multinomial Naive Bayes	0.7377	0.6910
Support Vector Machine	0.8500	0.6859
Gradient Booster	0.7074	0.6858
Logisitic Regression	0.7045	0.6857
Extreme Gradient Booster	0.7867	0.6830
Extra Trees Classifier	0.9996	0.6793
Adaptive Booster	0.6941	0.6744
Random Forest Classifier	0.9996	0.6692
K-Nearest Neighbors	0.6212	0.6114

Features

- Stemming
- Expanded Contractions
- No stop words
- CountVectorization

Final Model

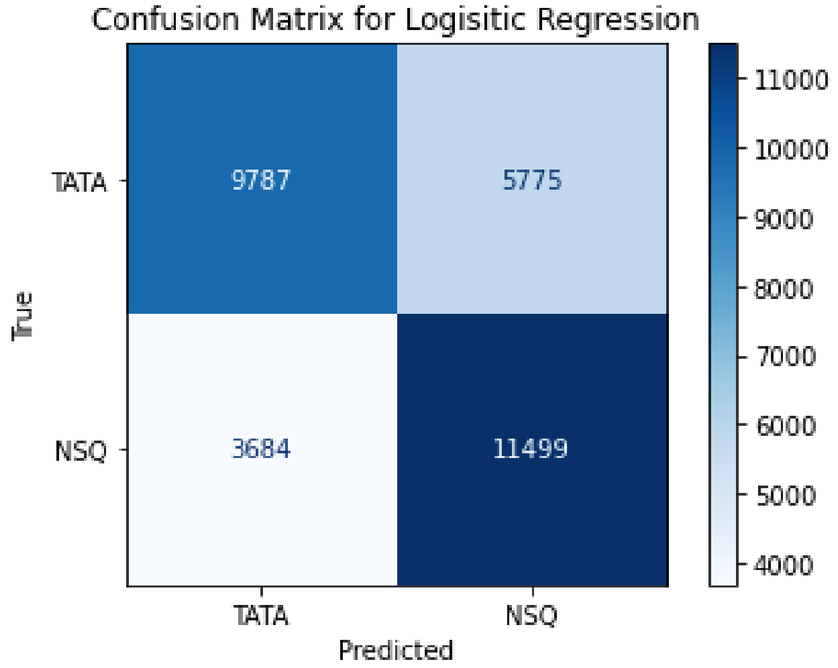
Logitistic Regression (122,978 posts)

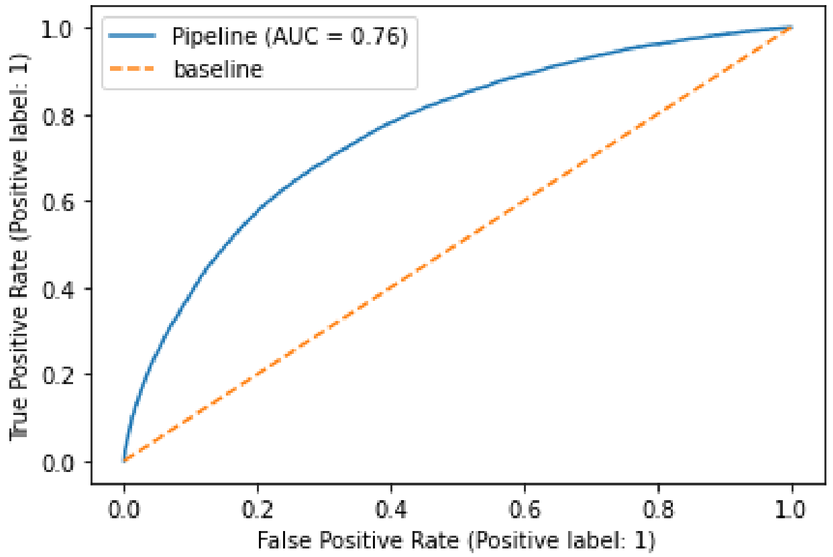
7 OUT 10 71.36%	
69.23%	

Analysis

- Misclassification: .3076
- Sensitivity: .7573
- Specificity: .6289
- Precision: .6657
- F1: .7086

- Model is better at classifying actual NSQ posts than actual TATA posts





Analysis

	Shared Words	NSQ only	TATA only
Word Count	15367	5153	4530
Percentage	61	21	18

NSQ Words

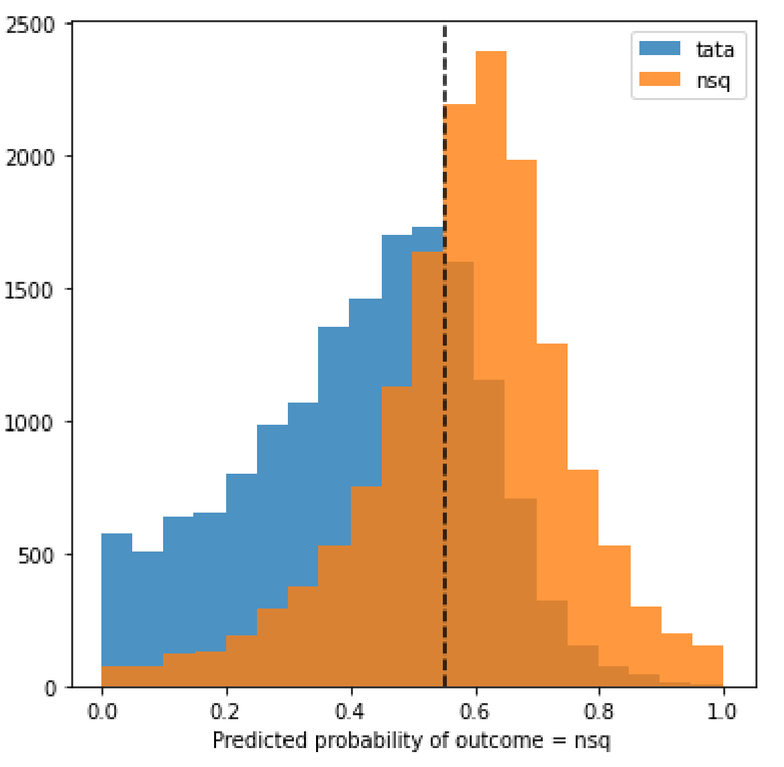
	Count
utm_medium	104
utm_sourc	86
ibb	68
ved	37
ios_app	36
iossmf	36
utm_nam	36
2c	35
android_app	34
wp	34

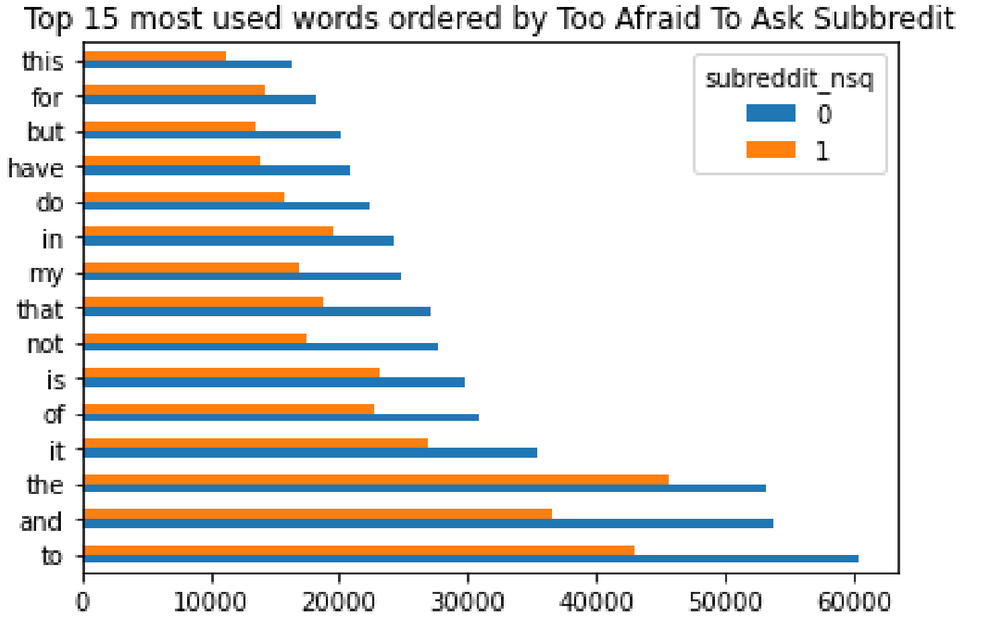
TATA Words

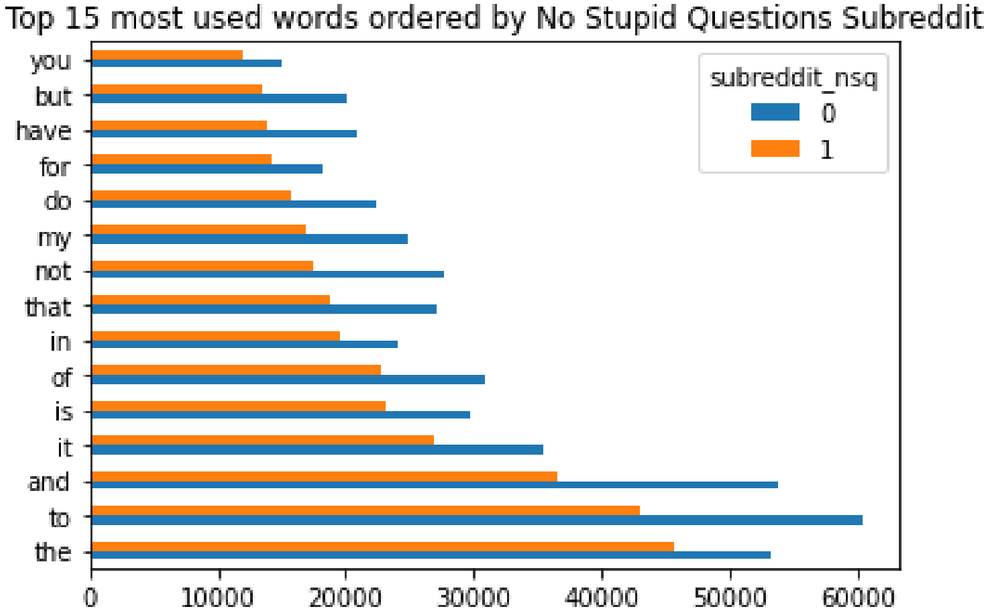
	Count
pedophilia	185
pedo	146
ivermectin	73
loli	66
uncircumcis	61
petito	49
mrna	44
fauci	42
euthanasia	41
castrat	38

- Large overlap of words between subreddits
- Low count for subreddit specific words

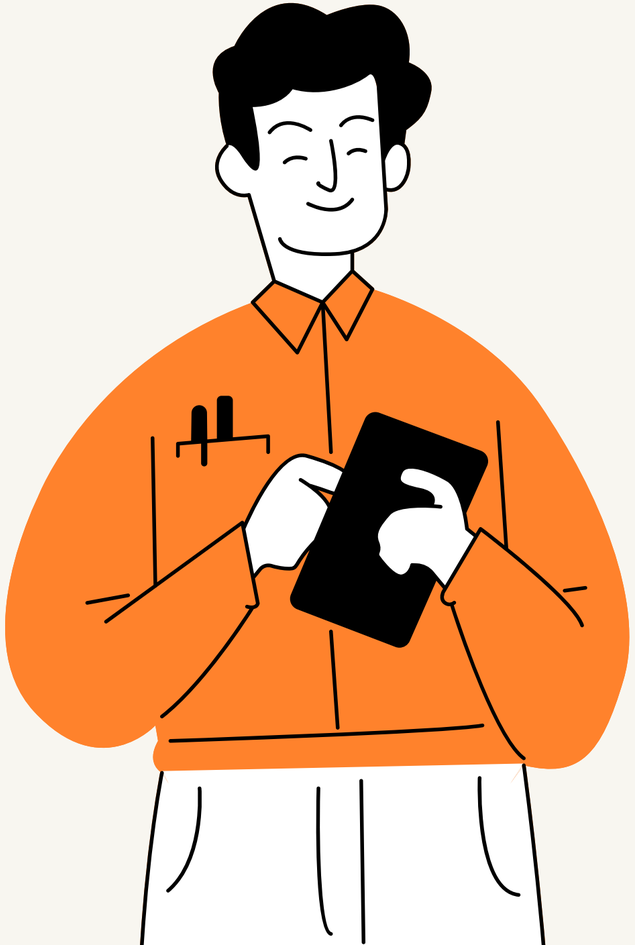
Results


tata
nsq
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability of outcome = nsq


Top 15 most used words ordered by Too Afraid To Ask Subbredit
subreddit_nsq
0
1
this
for
but
have
do
in
my
that
not
is
of
it
the
and
to
0
10000
20000
30000
40000
50000
60000


Top 15 most used words ordered by No Stupid Questions Subreddit
subreddit_nsq
0
1
you
but
have
for
do
my
not
that
in
of
is
it
and
to
the
0
10000
20000
30000
40000
50000
60000

Conclusion & Limitations

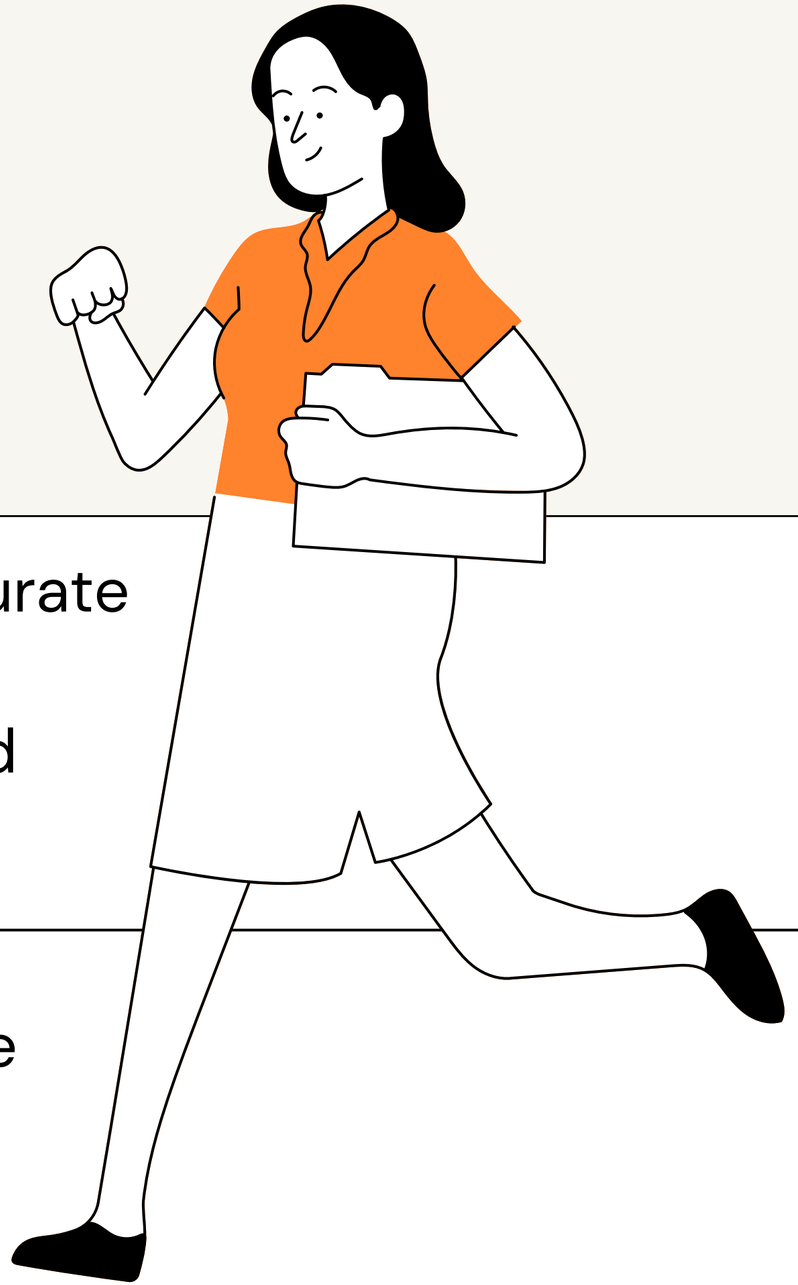


Could not produce accuracy scores above .8 using NLP; .69 or roughly 7 out of 10 correct classifications can be achieved.

Limitations due to the dataset and how the data was processed.

Optimizing the threshold, for classifying positives, has limited results (.005) due to overlap

Recommendations



- Alternative methods or features may be more accurate for this data set.
- Logistic regression based on created utc produced accuracy scores of .86 and .85 with less resources

- Highlight more specific topics/keywords i.e. include "link_flair_text" or "num_comments"