POP PUNK IS NOT PUNK

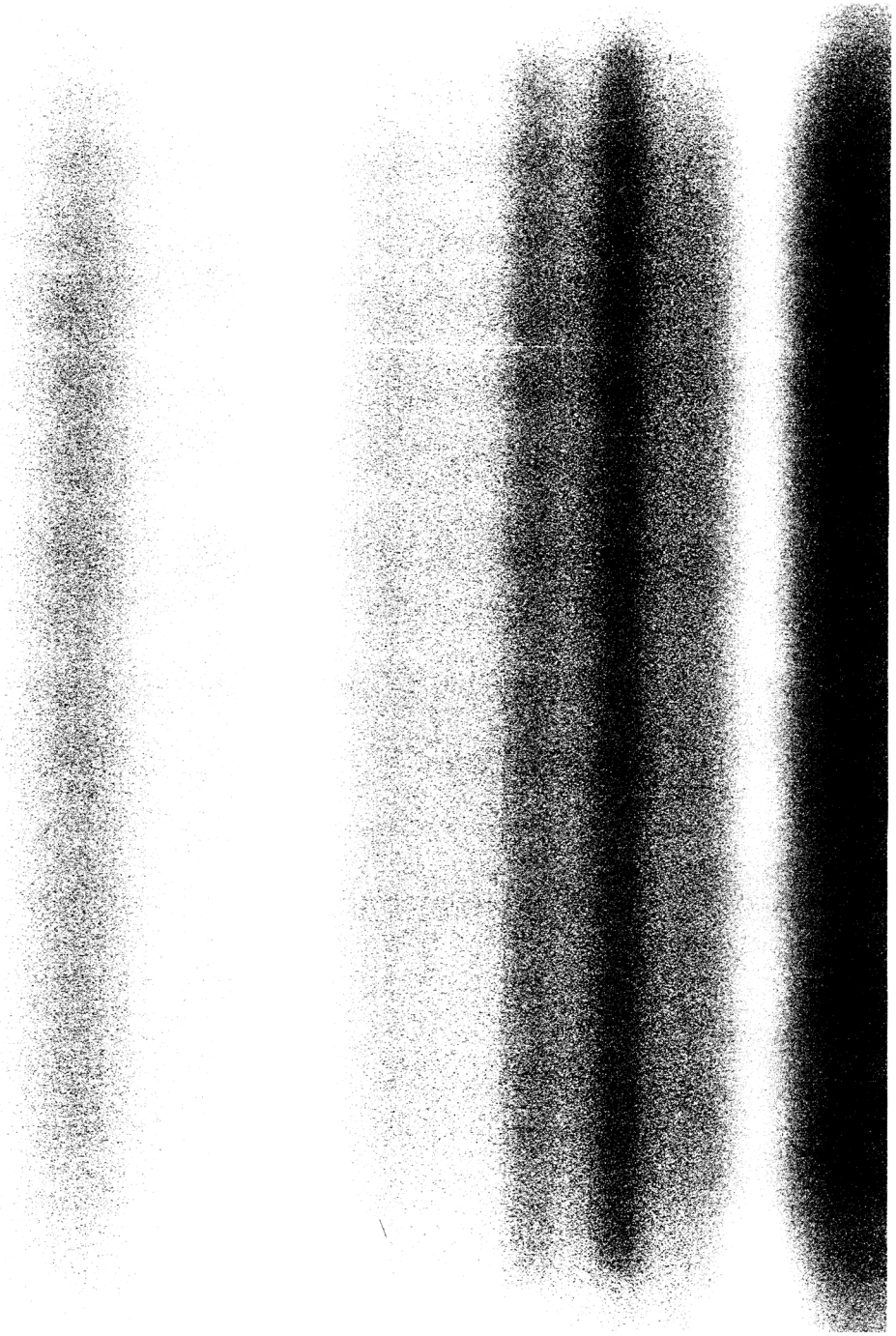
.....RIGHT?

PROBLEM STATEMENT

Is the language in text posts from the subreddits "Punk" and "PopPunkers" different enough that a classification model can predict which subreddit a post belongs to with an accuracy higher than the baseline?

OVERVIEW

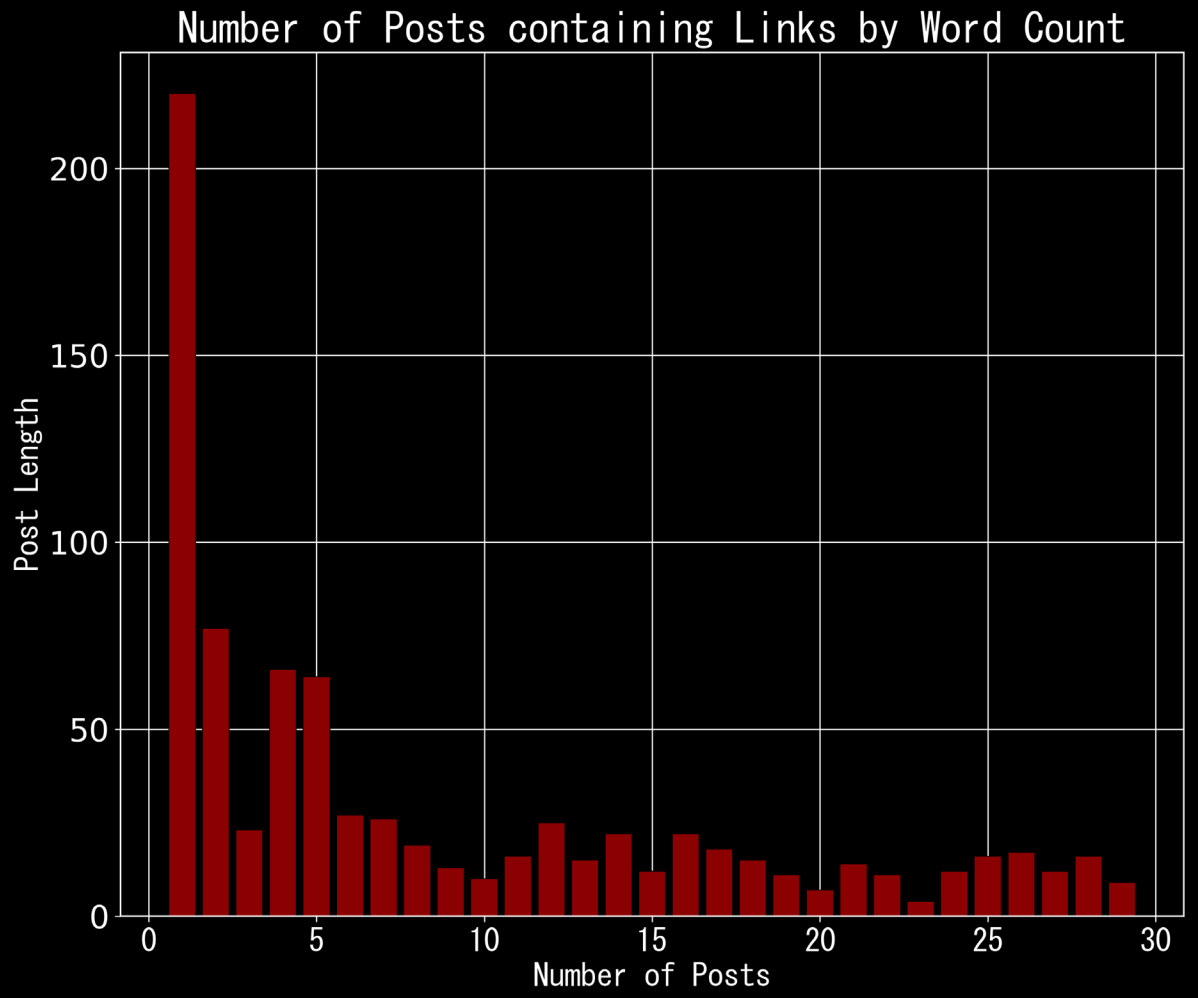
- What was our data?
- What model did we use?
- What did we find?
- How can we use this information?




OUR DATA

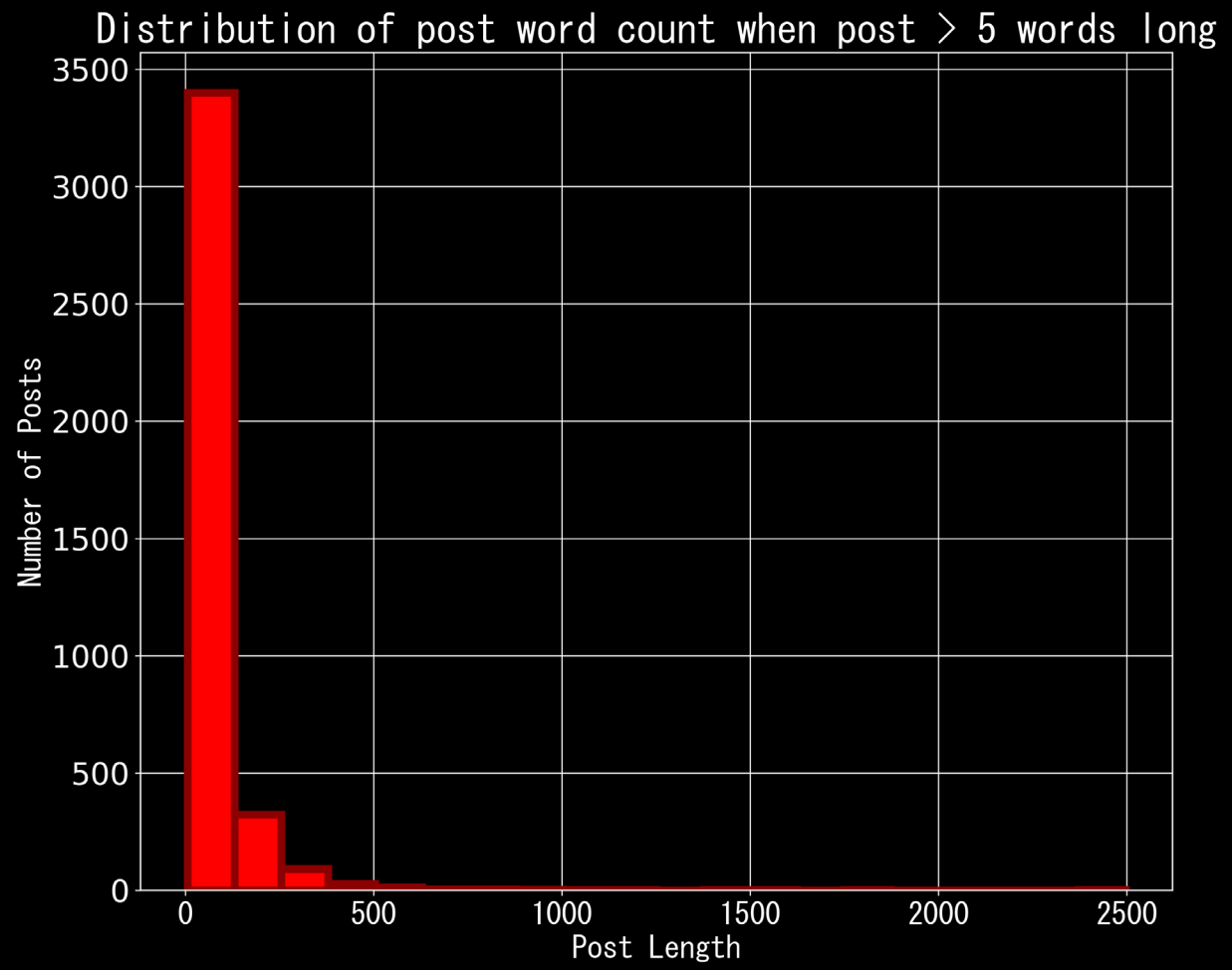
OUR DATA

- Scraped 5000 posts from Reddit
  - Pushshift API
  - 2500 from r/punk
  - 2500 from r/poppunkers
- A lot of single-word posts
  - Links
  - Deleted posts



Our Data

- Post length is skewed right



OUR DATA

5 most common words across both r/punk and r/poppunkers

WORD	COUNT
Punk	2972
Band	2161
Get	918
Pop	894
One	838

OUR MODEL

OUR MODEL

- Tested out three different models
  - Multinomial Naïve Bayes
  - Logistic Regression
  - Support Vector Machine
- Tested the models with two different vectorizers
  - Count Vectorizer
  - Tfidf Vectorizer

- Grid searched across them to find the best scores

GRIDSEARCH RESULTS

Test Number	model	cvec ngram range	tvec ngram range	train score	test score
1	LogisticRegression(max_iter=2000)	(1, 2)		0.788779	0.830721
3	LogisticRegression(max_iter=2000)	(1, 2)		0.787733	0.831766
5	LogisticRegression(max_iter=2000)	(1, 2)		0.787733	0.831766
2	MultinomialNB()	(1, 2)		0.812131	0.832811
4	MultinomialNB()	(1, 2)		0.812131	0.832811
0	MultinomialNB()	(1, 2)		0.811783	0.834901
6	SVC(degree=2, kernel='poly')		(1, 1)	0.793301	0.807732
7	SVC(degree=2, kernel='poly')		(1, 1)	0.793301	0.807732
8	SVC(degree=2, kernel='poly')		(1, 1)	0.793301	0.807732

*other params searched over included max df, and max features

OUR MODEL

- Count Vectorizer transformer with a Multinomial Naïve Bayes l classifier

INTERPRETABILITY

PERFORMANCE

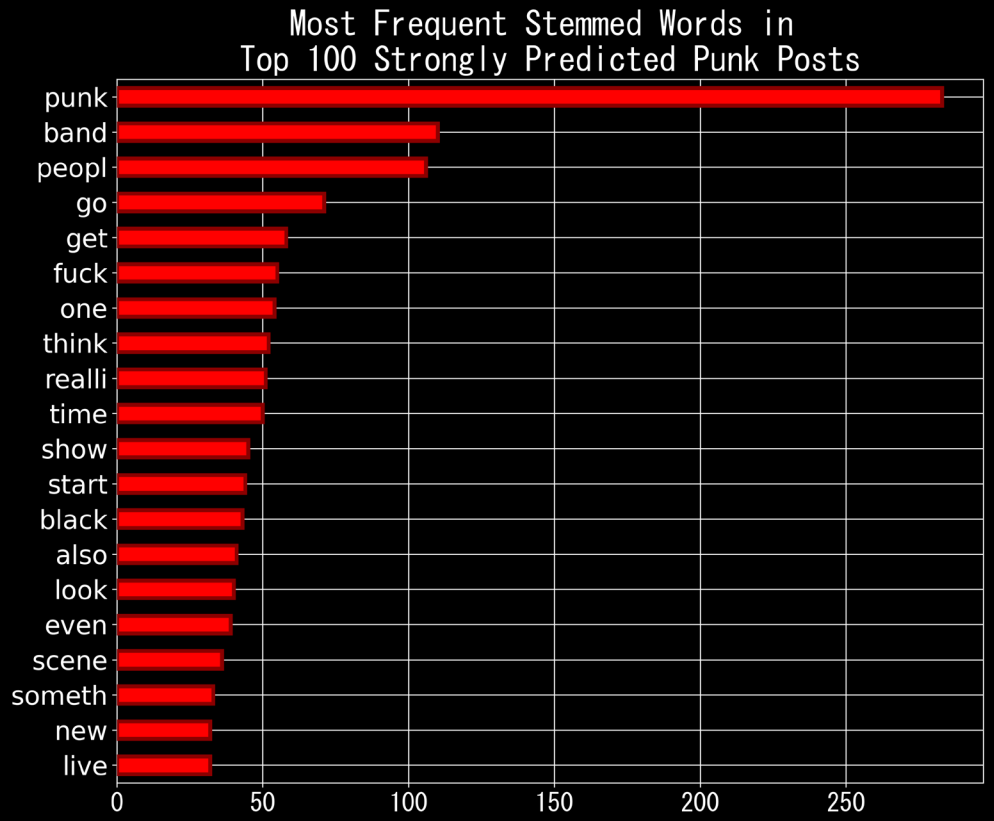
PREPROCESSING

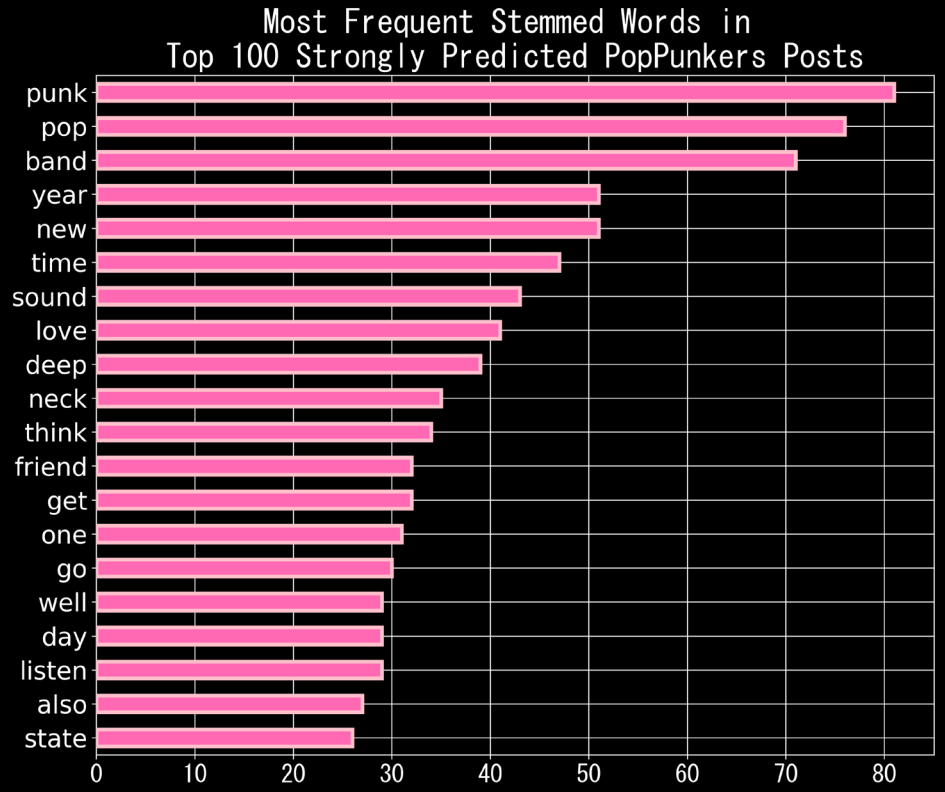
PERFORMANCE

- Baseline model: 50% accurate
- Cvec + Naïve Bayes Multinomial: 83.3% accurate on testing data

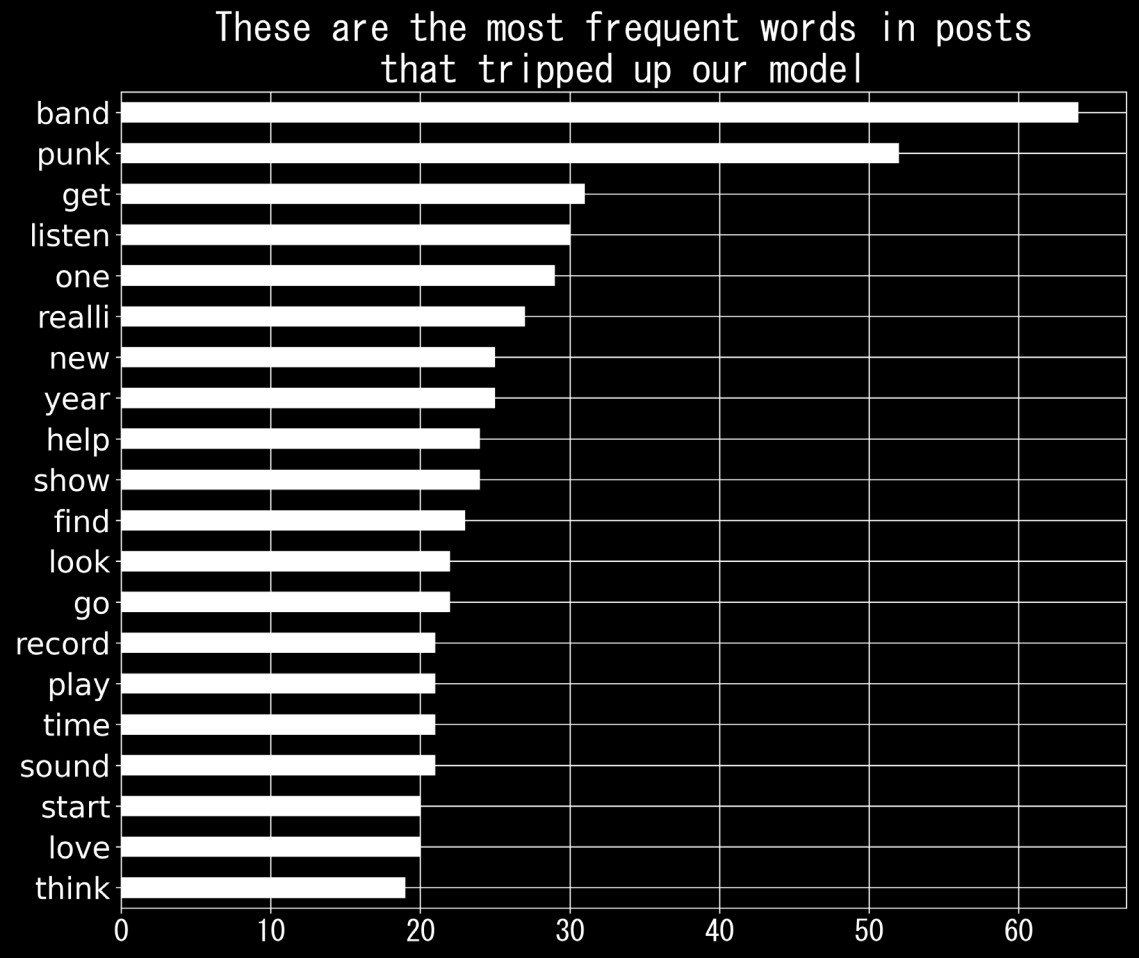
OUR FINDINGS

FINDINGS


Most Frequent Stemmed Words in
Top 100 Strongly Predicted Punk Posts
punk
band
peopl
go
get
fuck
one
think
realli
time
show
start
black
also
look
even
scene
someth
new
live
0
50
100
150
200
250


Most Frequent Stemmed Words in
Top 100 Strongly Predicted PopPunkers Posts
punk
pop
band
year
new
time
sound
love
deep
neck
think
friend
get
one
go
well
day
listen
also
state
0
10
20
30
40
50
60
70
80

FINDINGS


These are the most frequent words in posts
that tripped up our model
band
punk
get
listen
one
realli
new
year
help
show
find
look
go
record
play
time
sound
start
love
think
0
10
20
30
40
50
60

CONCLUSIONS

CONCLUSIONS

MODEL OUTPERFORMS BASELINE

SMALL OVERLAP BETWEEN SUBREDDITS

MARKETING AND MESSAGING


THANK YOU