

**POPPUNK IS
NOT PUNK**

.....RIGHT?

Presented by Eric Heidbreder for the
National Organization of Punk Rock Bands (NOPRoB)

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

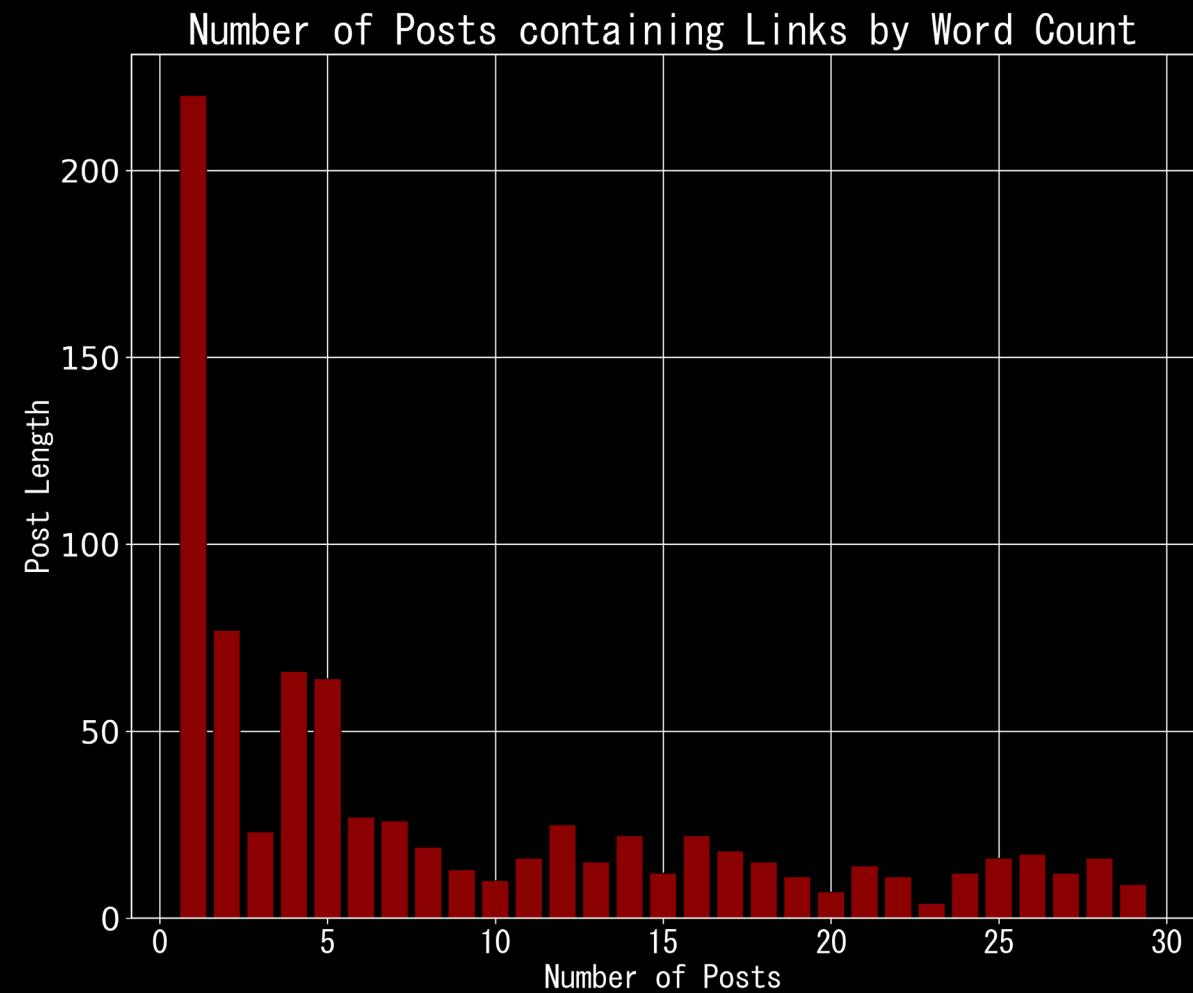
- What was our data?
- What model did we use?
- What did we find?
- How can we use this information?



OUR DATA

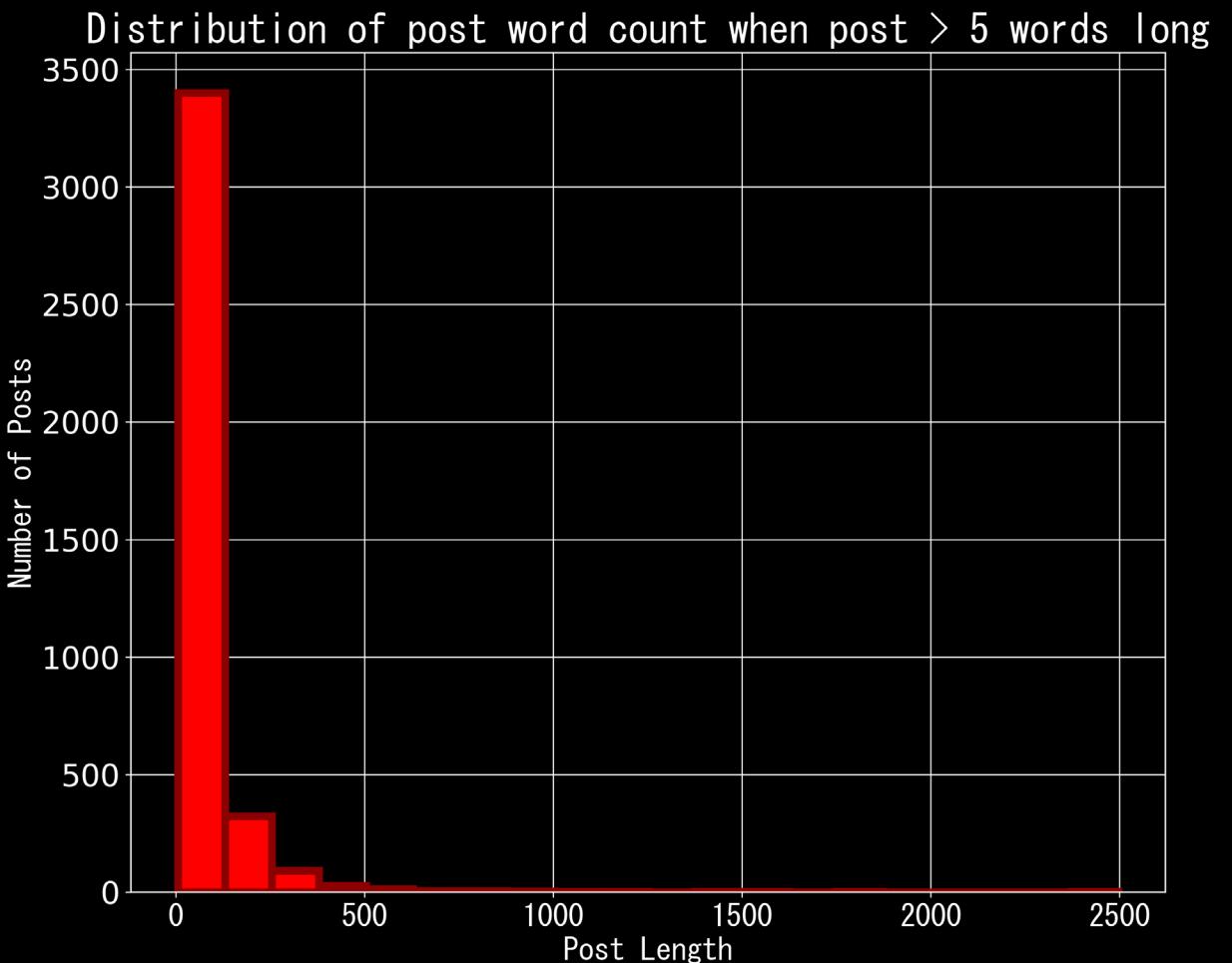
OUR DATA

- Scraped 5000 posts from Reddit
 - Pushshift API – selftext posts only
 - 2500 from r/punk
 - 2500 from r/poppunkers
- A lot of single-word posts
 - Links
 - Deleted posts
- Focused on posts with a word count over 5
 - 1961 from r/punk
 - 1916 from r/poppunkers



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
 - Tfifd 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	(1, 2)		0.788779	0.830721
3	LogisticRegression	(1, 2)		0.787733	0.831766
5	LogisticRegression	(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		(1, 1)	0.793301	0.807732
7	SVC		(1, 1)	0.793301	0.807732
8	SVC		(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 classifier

INTERPRETABILITY

PERFORMANCE

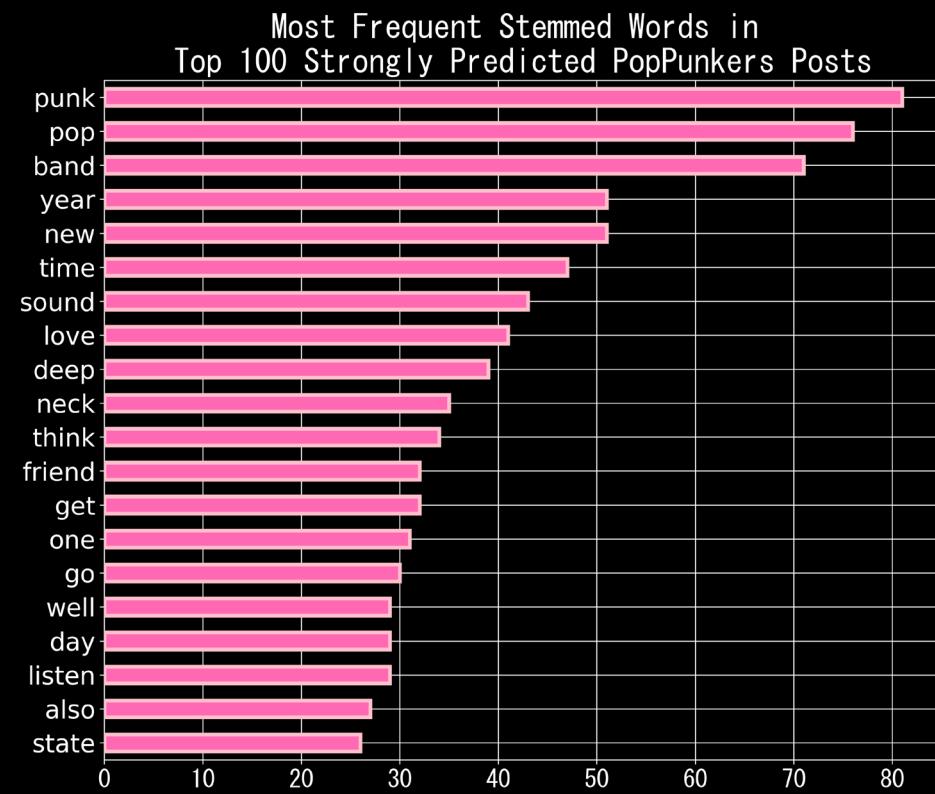
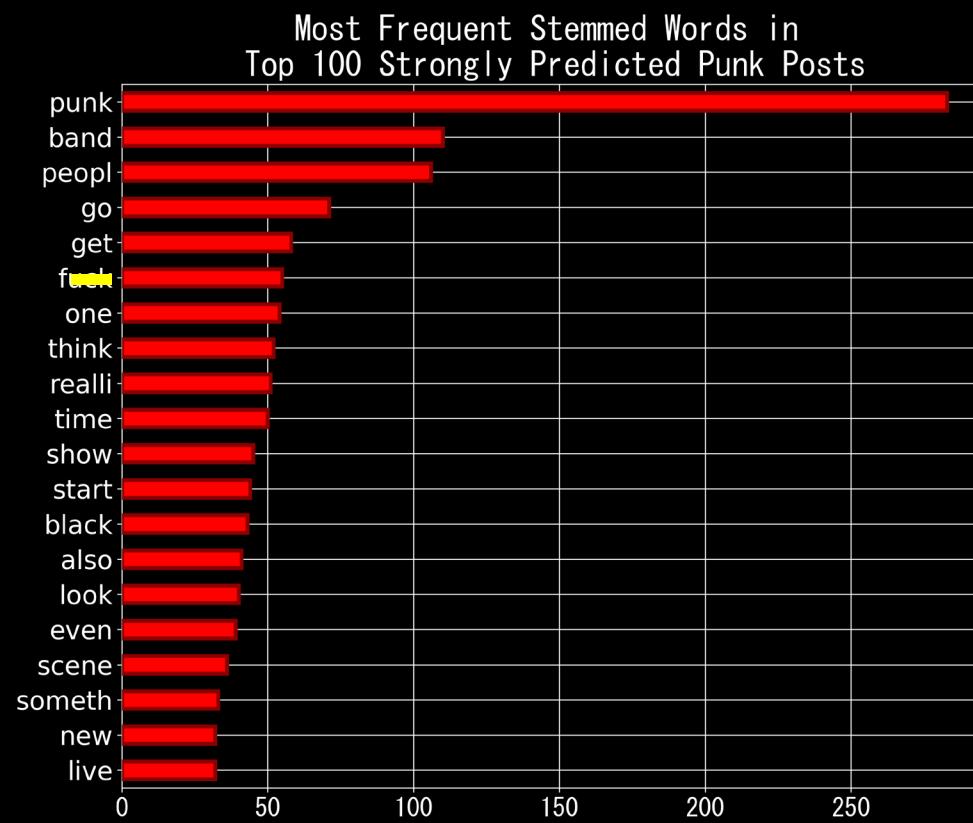
PREPROCESSING

PERFORMANCE

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

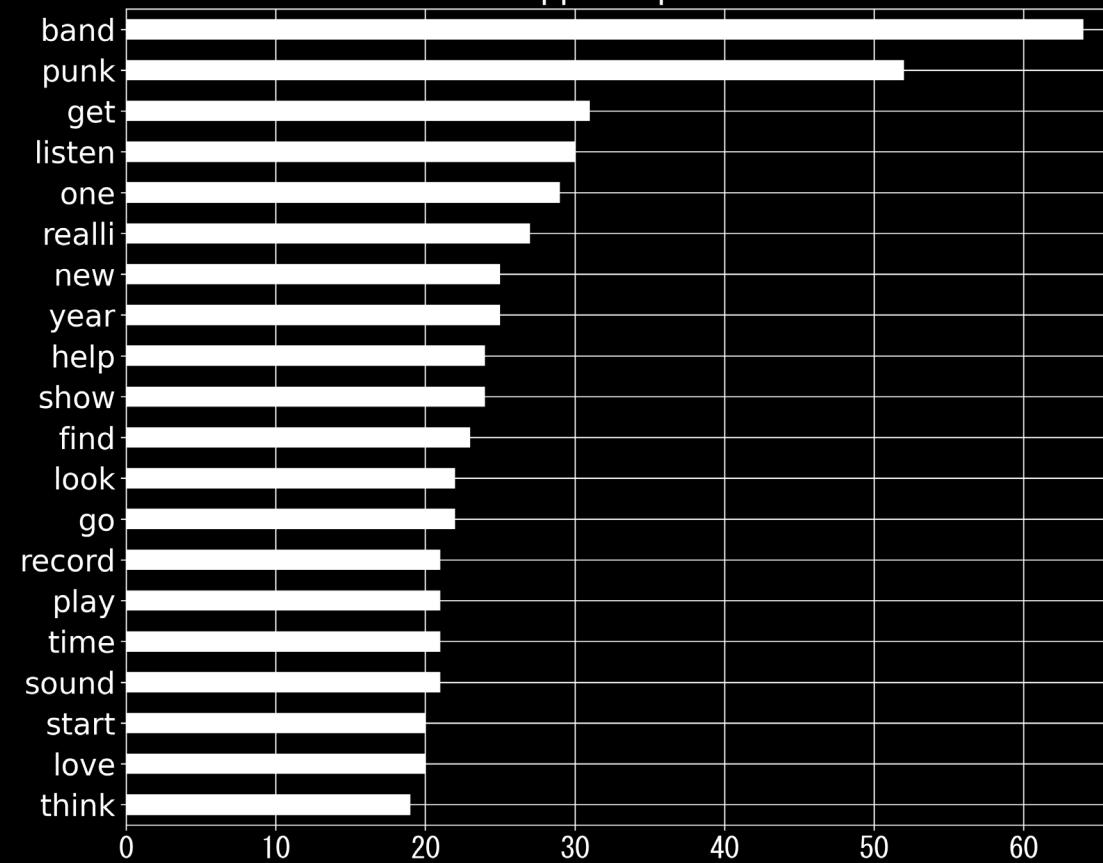
OUR FINDINGS

FINDINGS



FINDINGS

These are the most frequent words in posts
that tripped up our model



CONCLUSIONS

CONCLUSIONS

MODEL OUTPERFORMS BASELINE

SMALL OVERLAP BETWEEN SUBREDDITS

MARKETING AND MESSAGING

SOURCES

- api.pushshift.com
- reddit.com/r/punk/
- reddit.com/r/poppunkers/

THANK YOU