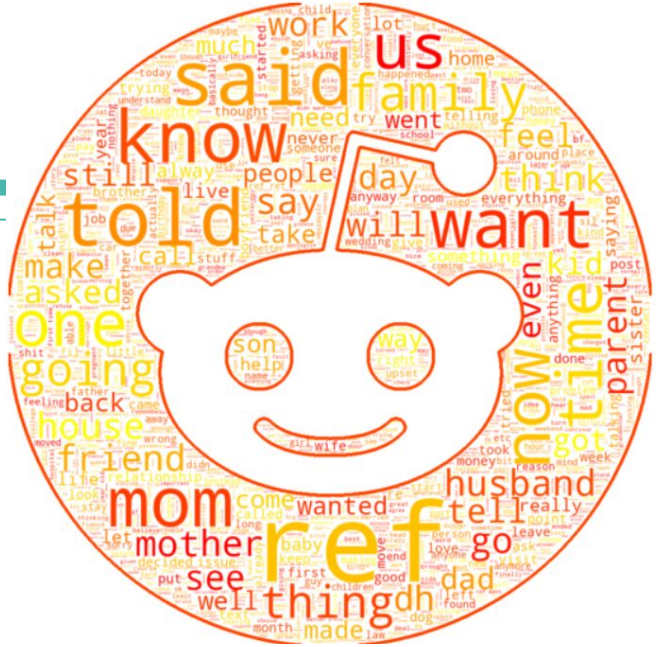



# Reddit Subreddit Sorting

**Arink Bertrand**
Fellow at General Assembly - DSI Course

# Can We Identify subreddits based on 1-3 word combinations?

# The Subreddits



2

3

4

# The Data

5

6

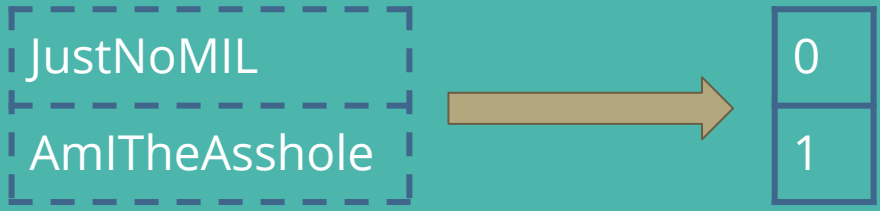
7

- Scrapped Reddit using PushShift API
- 9000 posts from r/AmITheAsshole
- 7000 posts from r/JustNoMIL

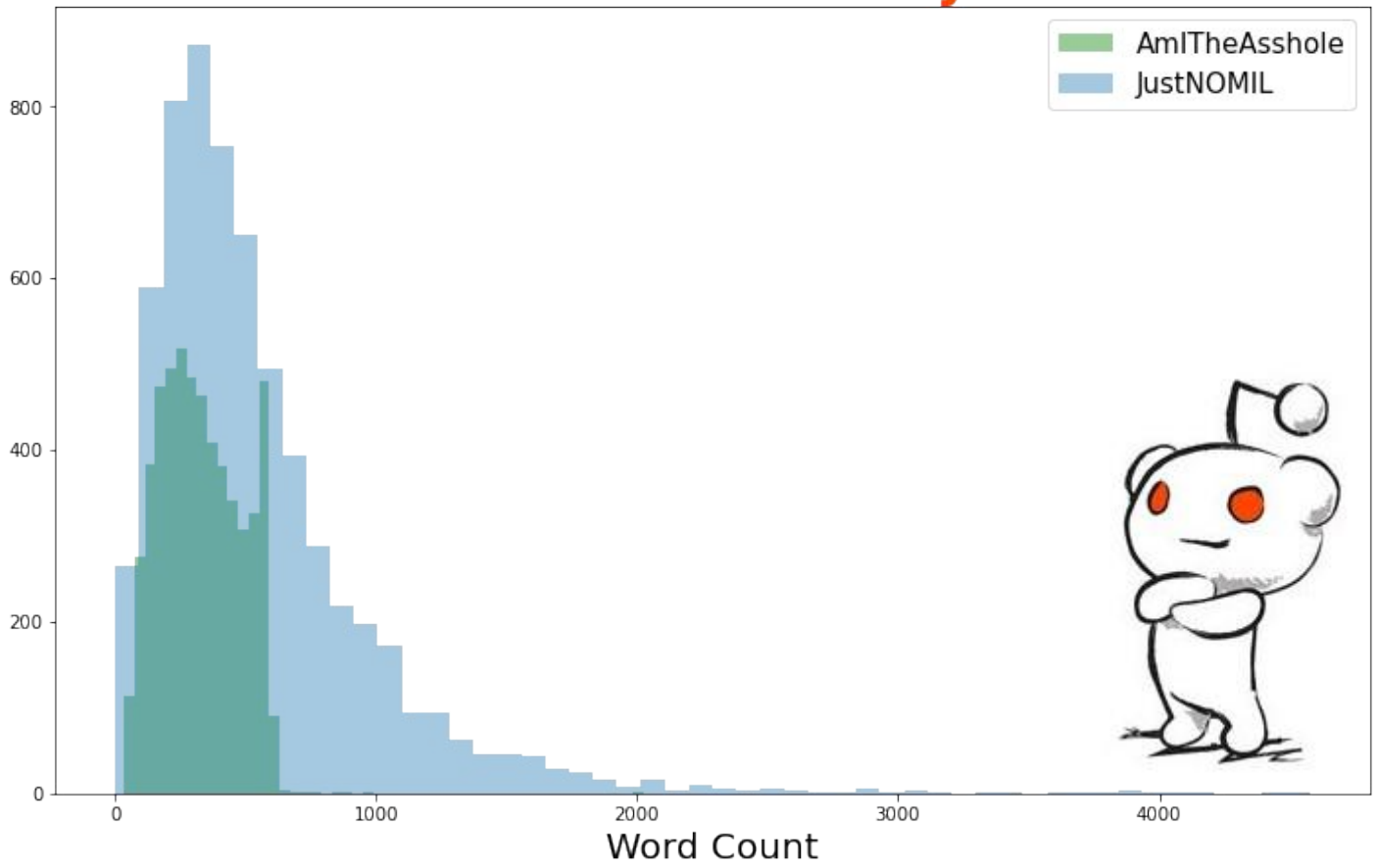
8

# Data Cleaning

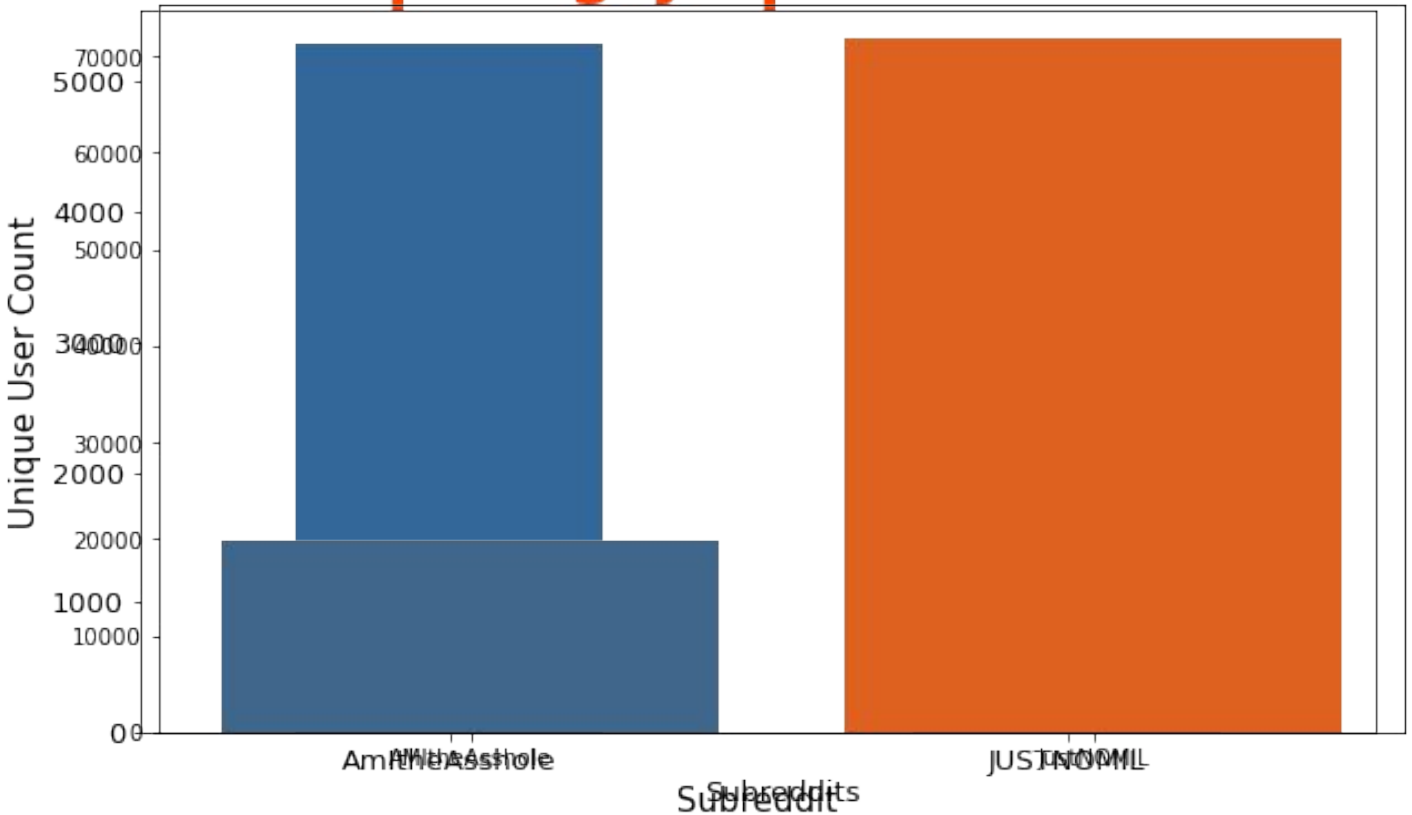
- [deleted] & [removed]
  - Duplicates
- AutoModerator posts
  - \n
  - 




# Word Count Distribution by Subreddit



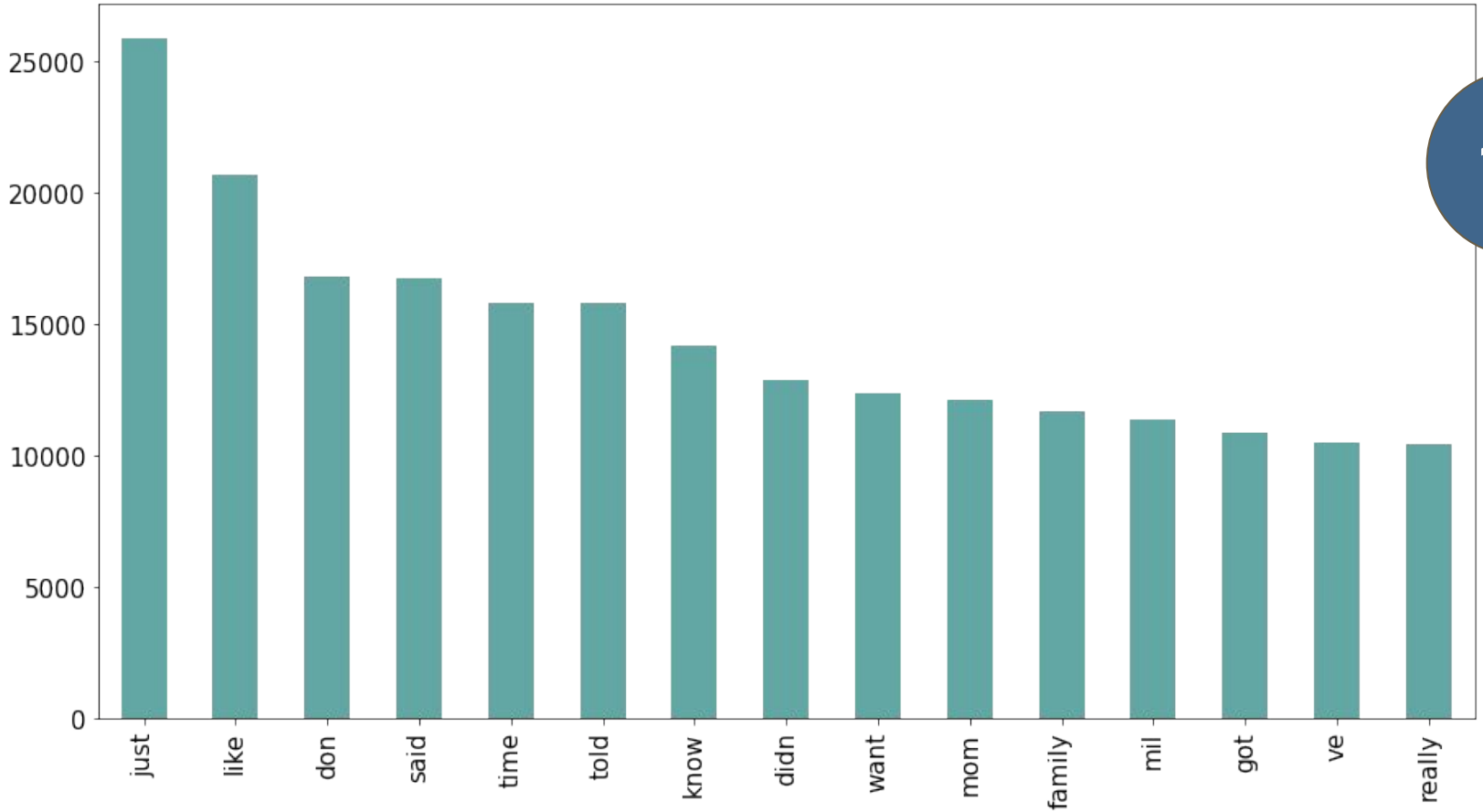
# How Angry Are The Users?




11

10

# Most Common Words in Titles & Posts



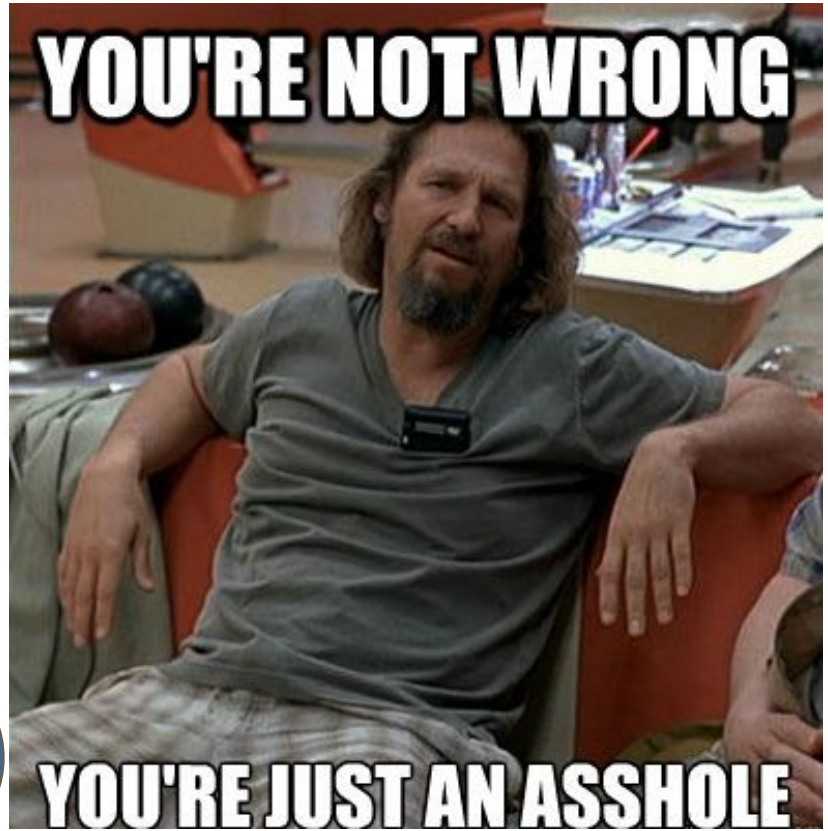
# Model Prep

- StopWords = ‘english’ + ‘top 15’
- ‘ref’  for identifiers
- Combined title and selftext columns
- Target = ‘sub-reddit’
- Train / Test split

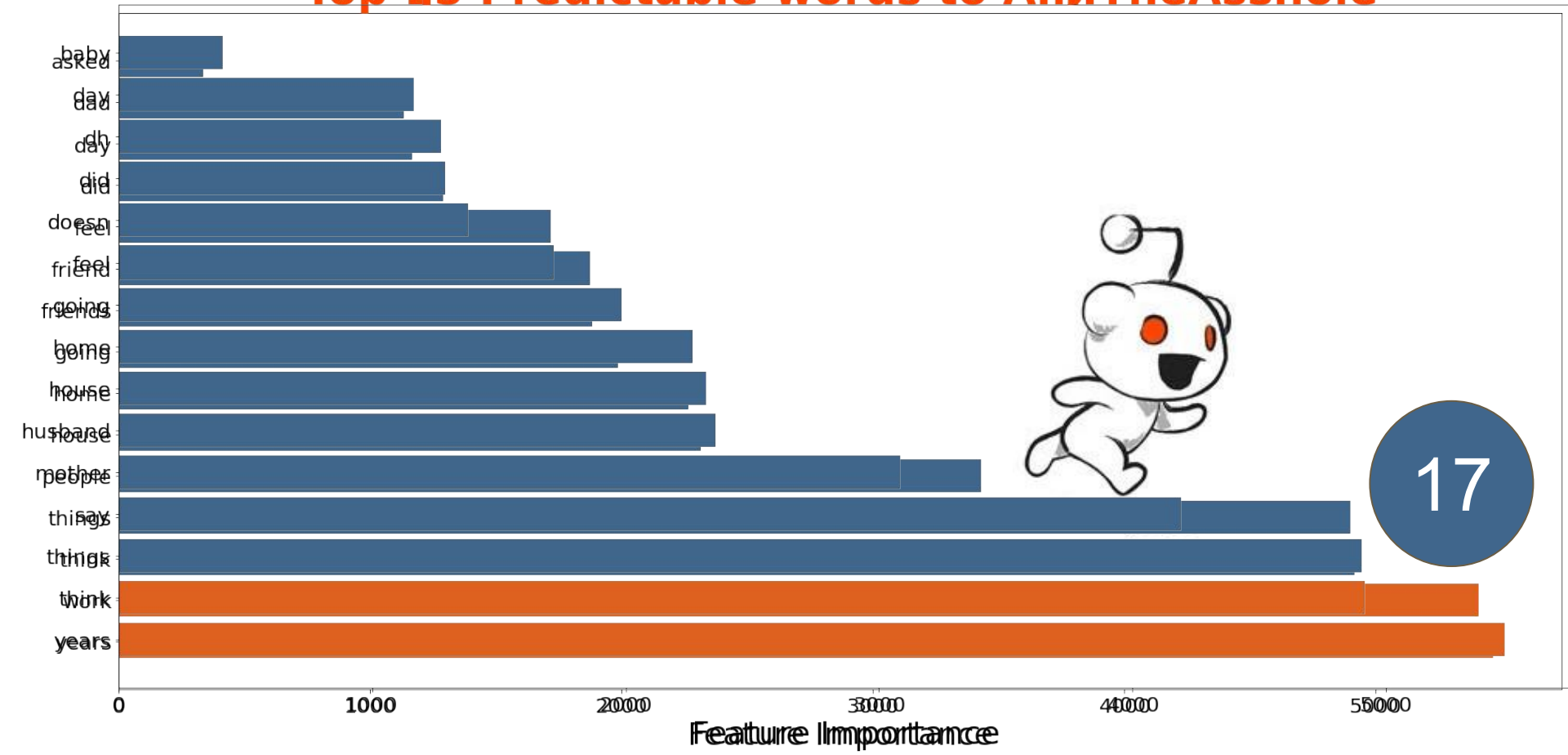
# Models w. GridSearch Pipeline

- Transformer
  - CountVectorizer
  - TFIDFVectorizer
- Estimator
  - MultinomialNB
  - BernoulliNB
  - LogisticRegression
  - AdaBoost Classifier

16 15 14 13

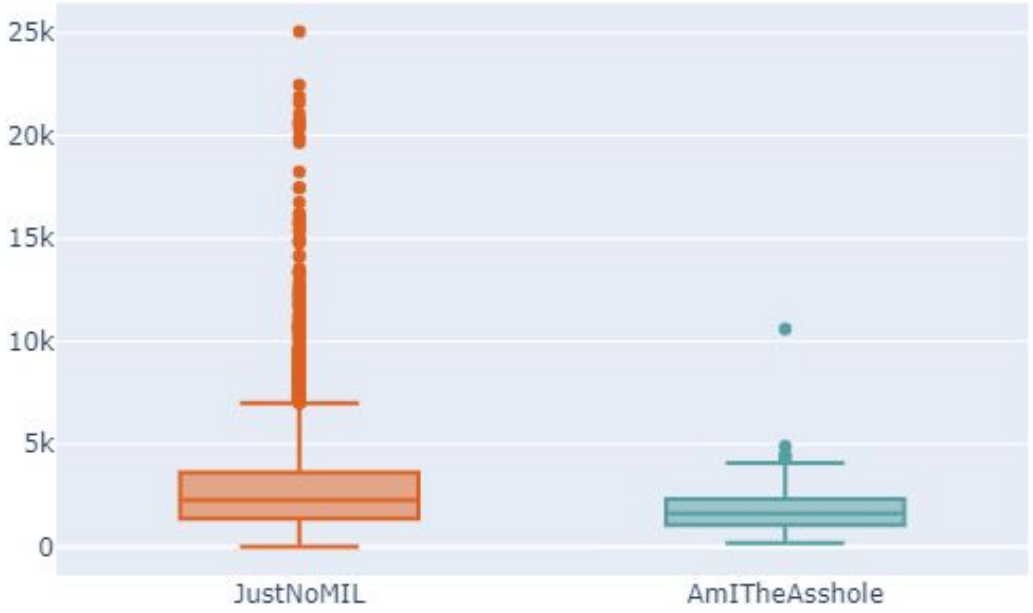


# Top 15 Predictable words to JustNOMIL



| | | lass. |
|---|---|---|
| | | Vect |
| Train | | 9.51% |
| Test | | 7.59% |
| Specif | | 6.49% |

# Post Length per Subreddit




# THANK YOU!