# NLP Classification Project

by Marta Lew

# Problem Statement

Build and algorithm which will accurately classify short texts based on whether its author is a man or a woman*.

**Some possible real-life applications:**

- Data enrichment (e.g. medical survey entries with missing information)
- Prevention of social engineering (catfishing)
- Digital marketing

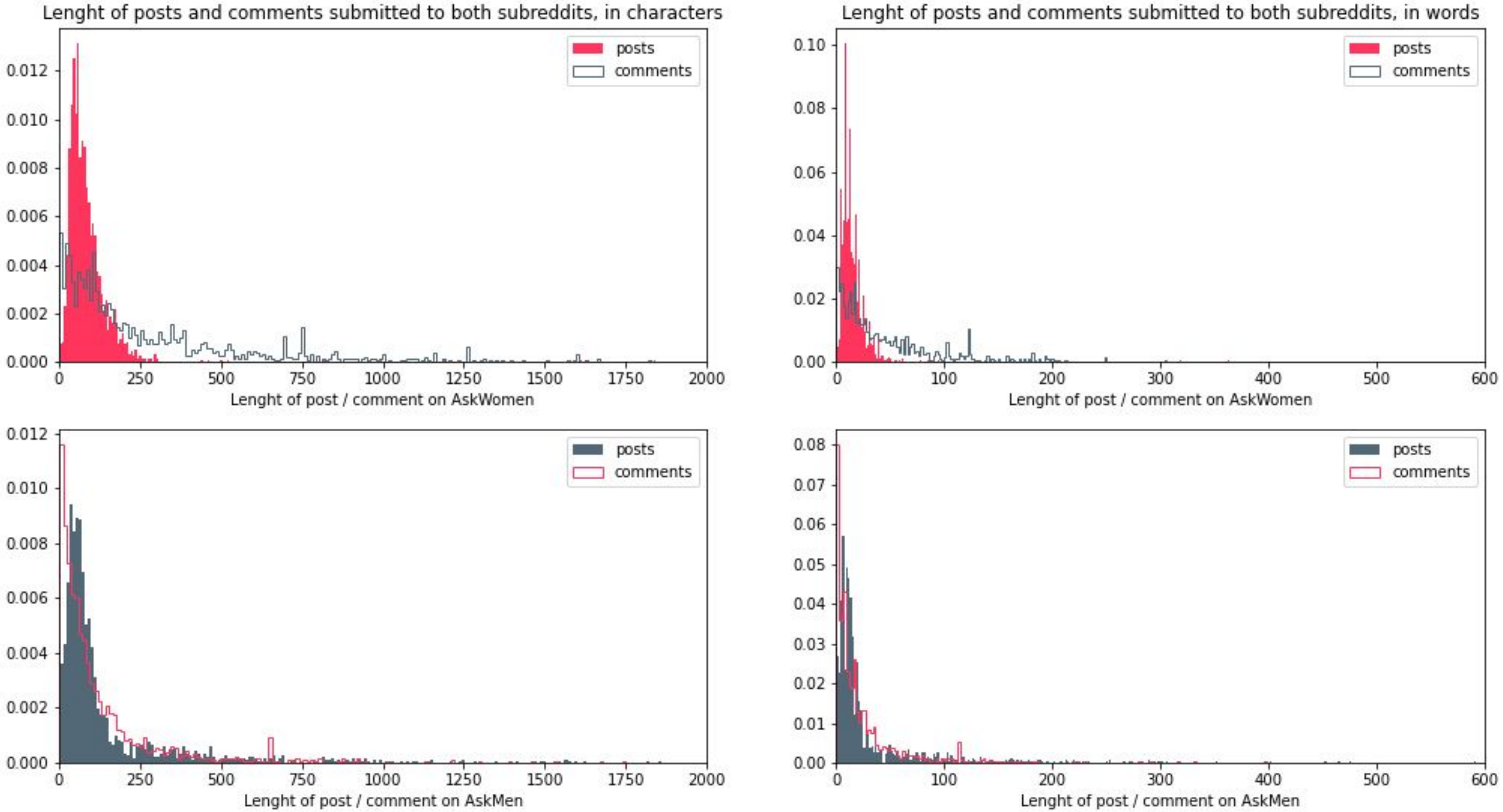
'* While gender is known to not be a binary construct, this project is set up as a binary classification

# The dataset

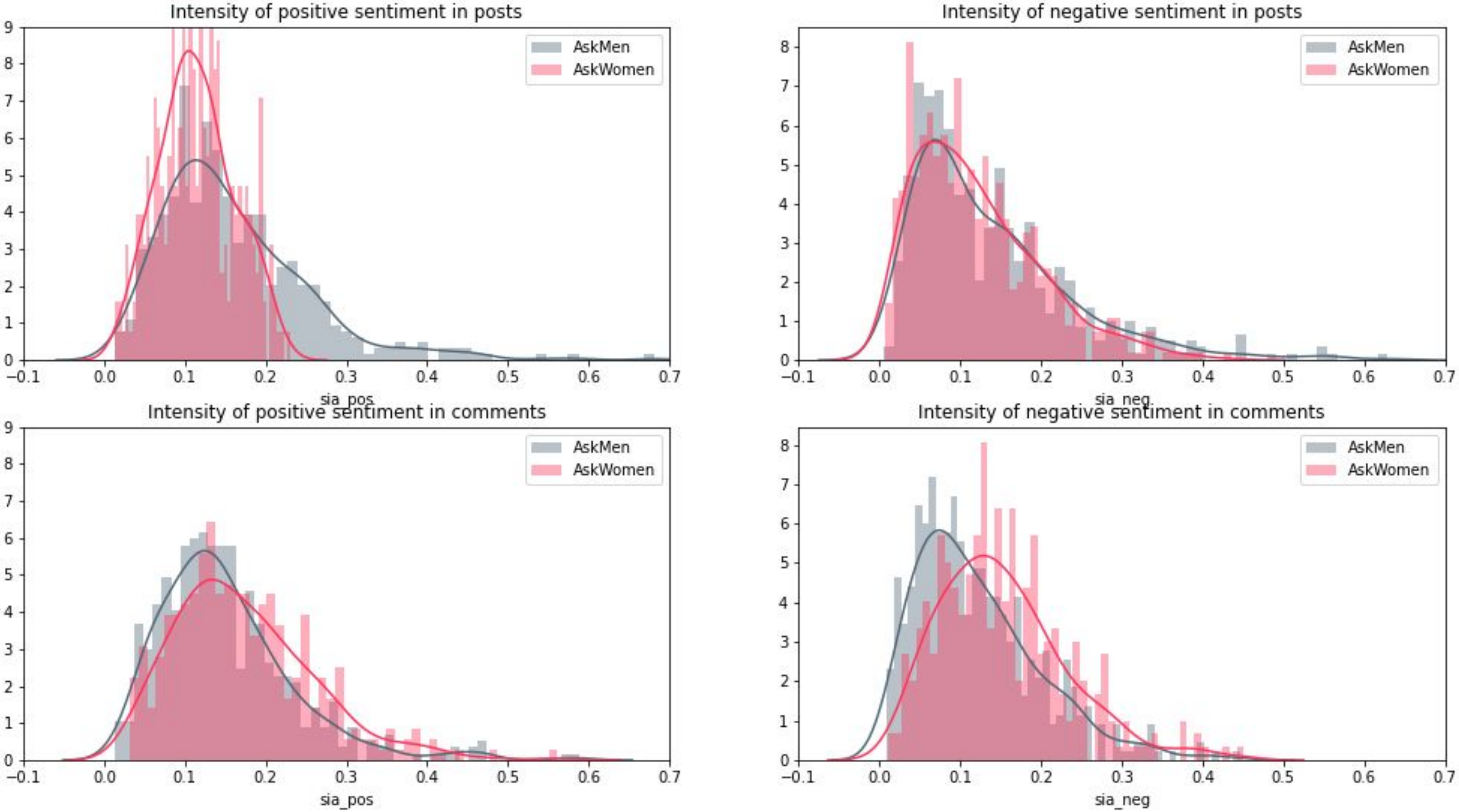
| Subreddit | posts | comments |
|---|---|---|
| AskWomen | 3,000 | 2,685 |
| AskMen | 2,900 | 2,944 |
| **Total** | **5,900** | **5,629** |
| | **11,529** | |

- Posts and comments were pulled using the Pushfit API
- All posts and comments were entered in a 7-day period from 12/21 to 12/28/2020
- Analyzed text did not include comments made by Reddit moderators
- It was assumed that all other text on AskWomen was authored by women, and AskMen - by men

# Length of posts and comments, by Subreddit



# Sentiment intensity, by Subreddit



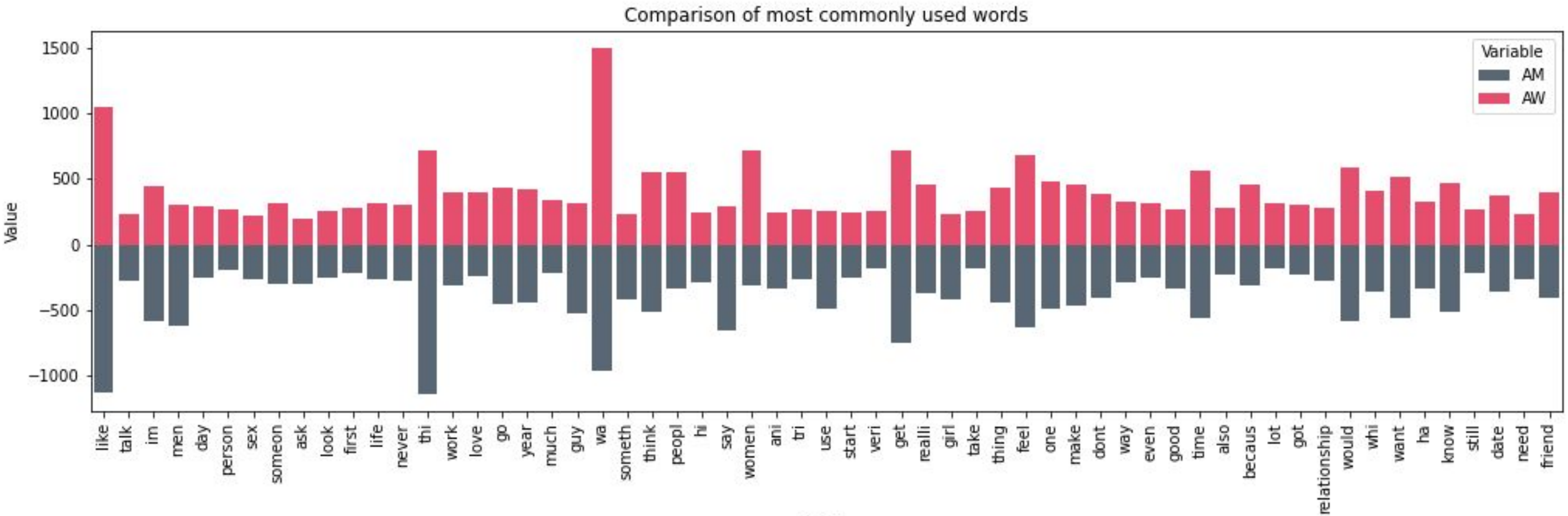
Intensity of positive sentiment in posts

Intensity of negative sentiment in posts

Intensity of positive sentiment in comments

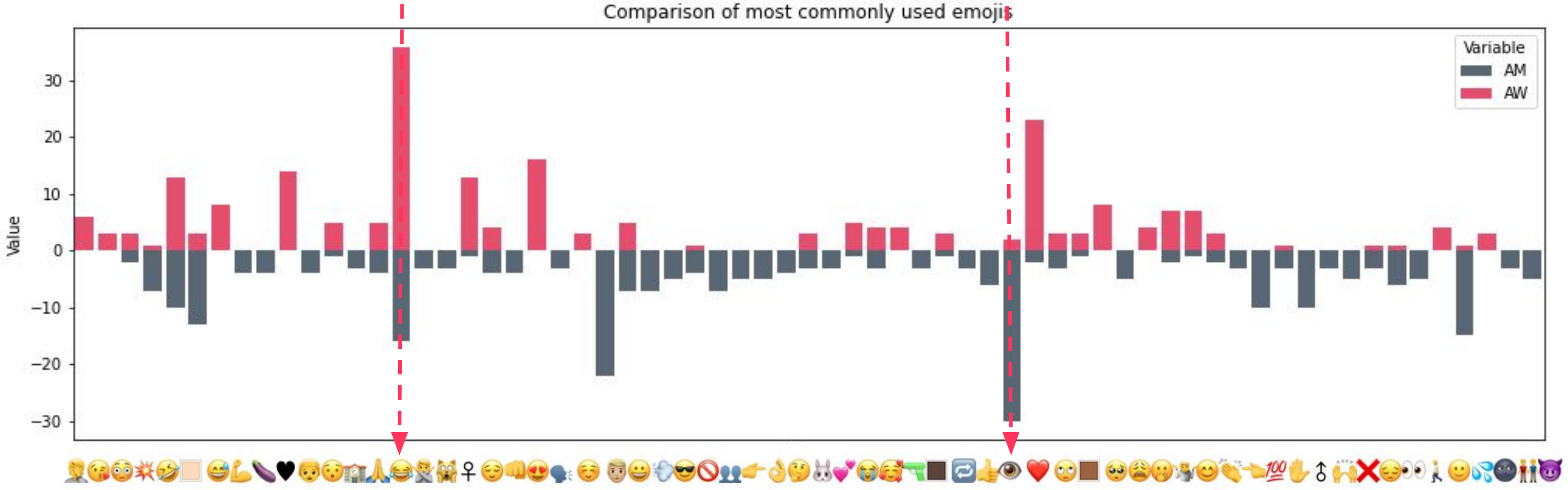
Intensity of negative sentiment in comments

# Most common words in each Subreddits



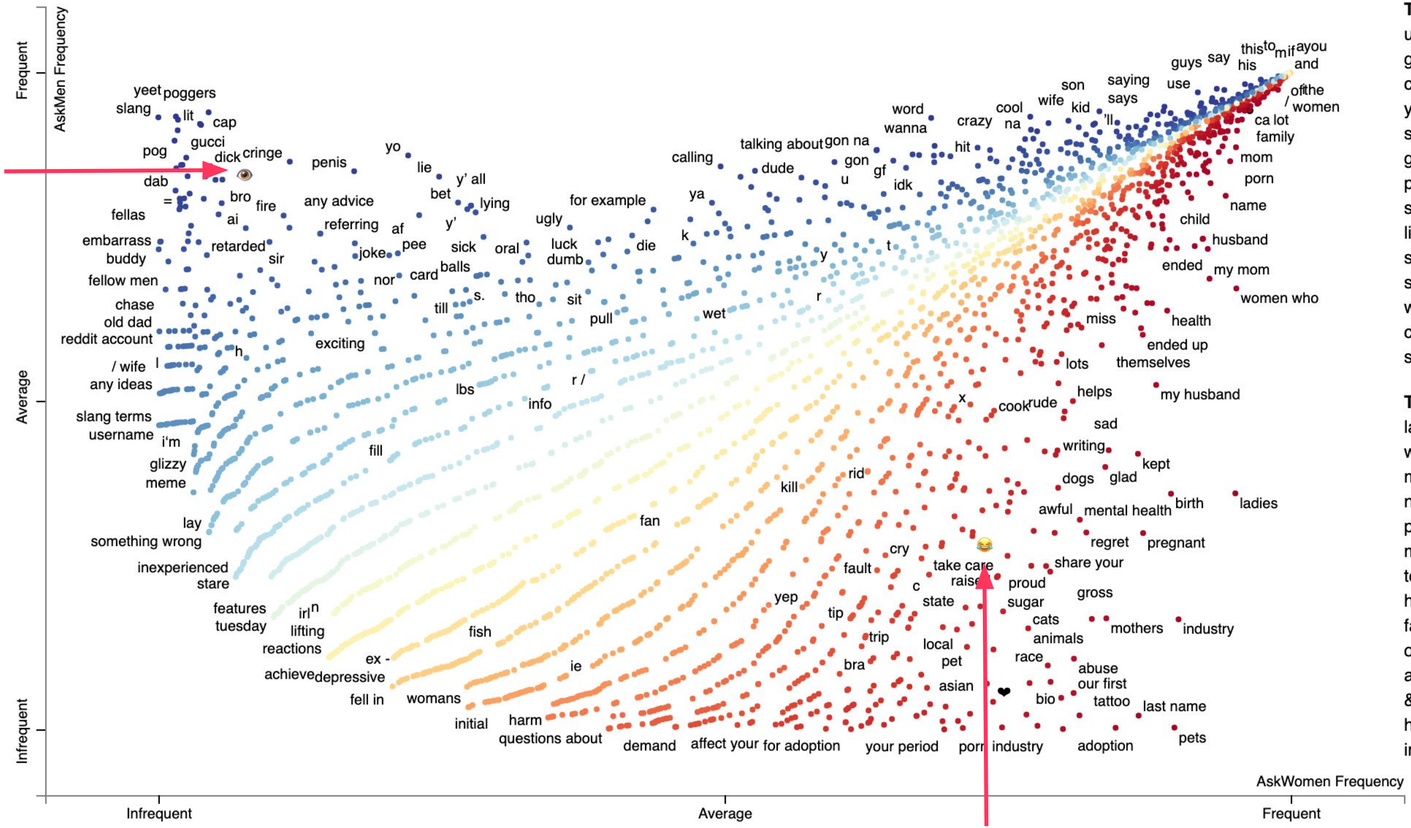
- Of the top 50 words used by men and women respectively, 59 (shown above) were shared
- Words frequently used by men and not used much by women: 'ani', 'ask', 'girl', 'hi', 'need', 'sex', 'someth', 'start', 'talk'
- Words frequently used by women and not used much by men: 'also', 'first', 'got', 'lot', 'much', 'person', 'still', 'take', 'veri'

# Most commonly used emoji, by Subreddit



- Men used 175 different emoji and used them 439 times women - only 67 emoji, 271 times
- Emojis were more commonly used in comments (630 uses) than posts (80 uses)
- On average, men used 0.14 emoji per comment, women - only 0.08
- The comment with most emoji had 343(!) of them and appeared on AskMen

# A different view of frequent terms



**Top AskMen**

use
guys
cap
yeet
slang
girl
poggers
sus
lit
son
say
word
cool
saying

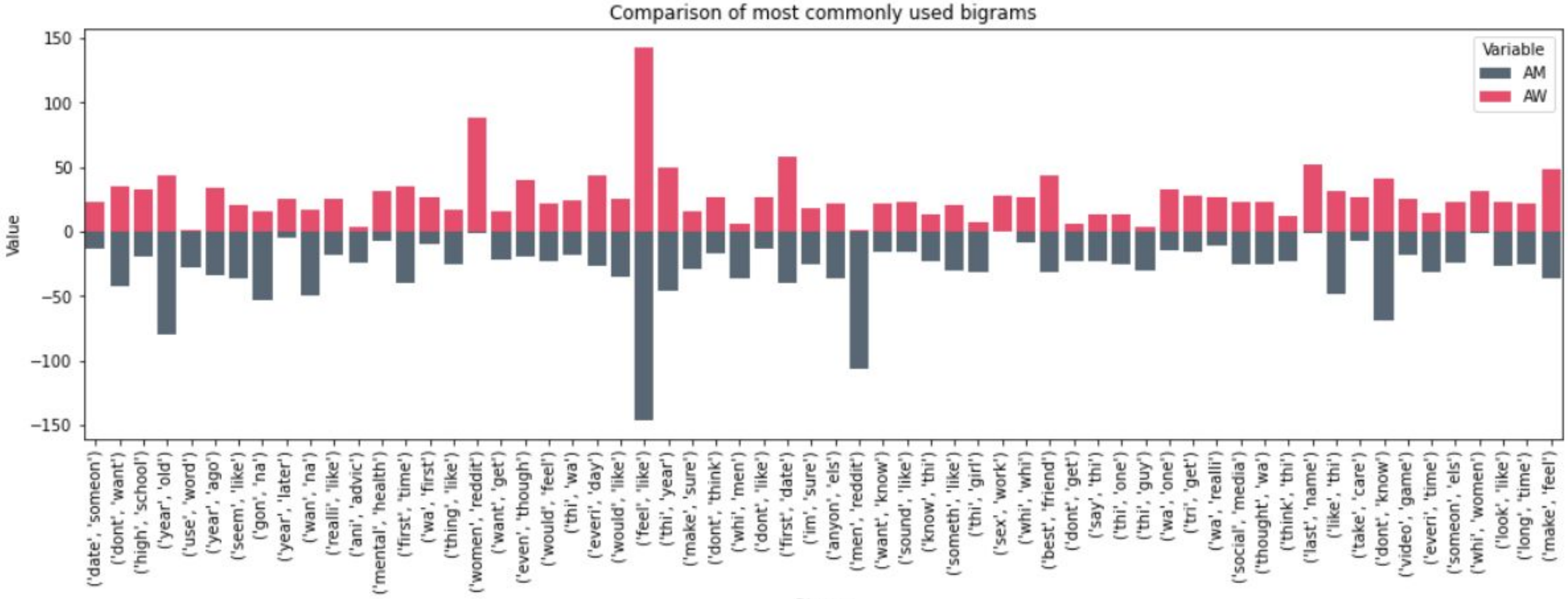
**Top AskWomen**

ladies
women who
mom
name
porn
my mom
together
husband
family
our
amp
& amp
hope
industry

**Characteristic**

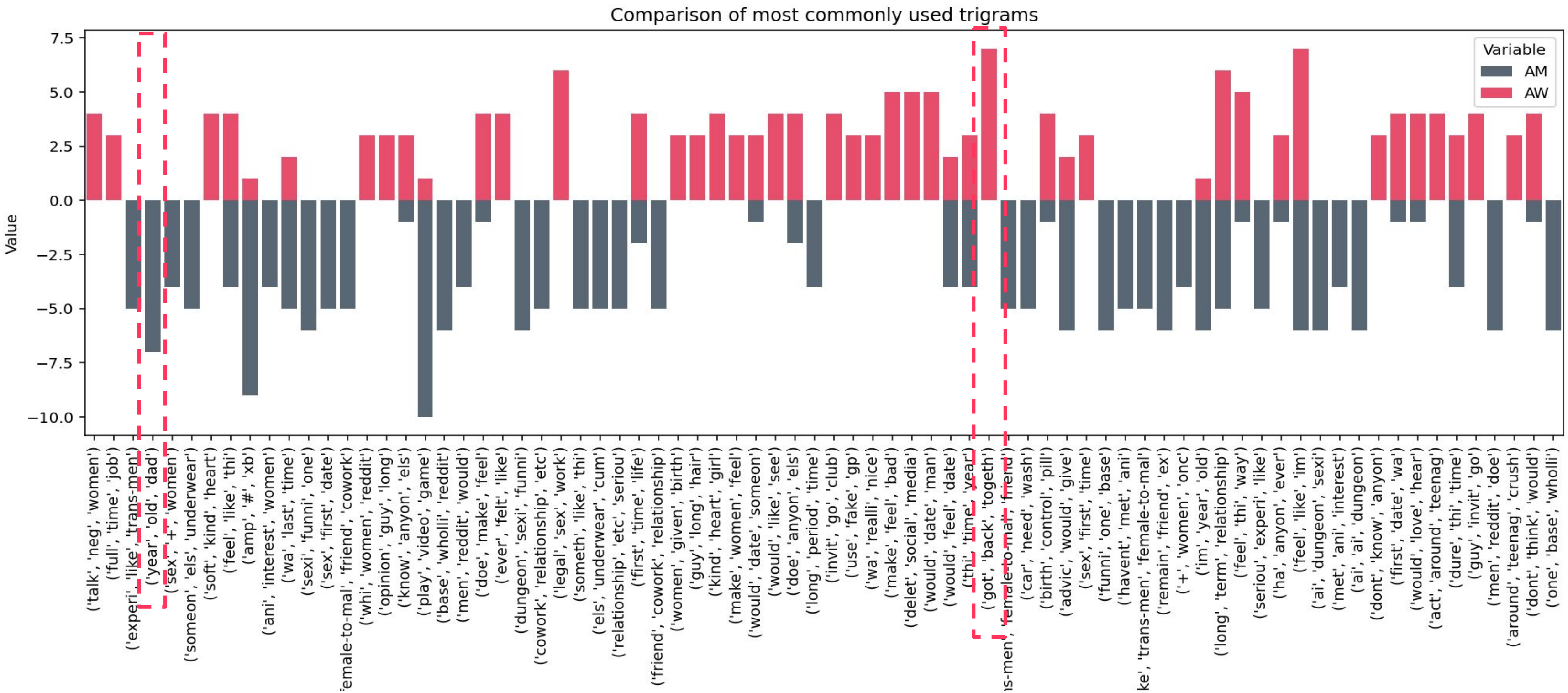
reddit
yeet
covid
poggers
pog
bruh
fleek
instagram
texting
idk
pogchamp
deadass
tinder
texted
slaps
cringe
nt
mwy
gon
onlyfans
insecurities
snapchat
wouwd
slang
fortnite
youtube
coworker
hiws
facetime
nsfw

# Most common bigrams, by Subreddit



- 62 of the top 40 bigrams used by men and top 40 used by women were common to both Subreddits
- Top 2-word phrases  frequently used by men and not by women: (‘men’, ‘reddit’), ('use', 'word'), ('thi', 'girl')
- Top bigrams  frequently used by women and not by men:  ('women', 'reddit'), (‘last’, ‘name’), (‘sex’, ‘work’)

# Most common trigrams, by Subreddit

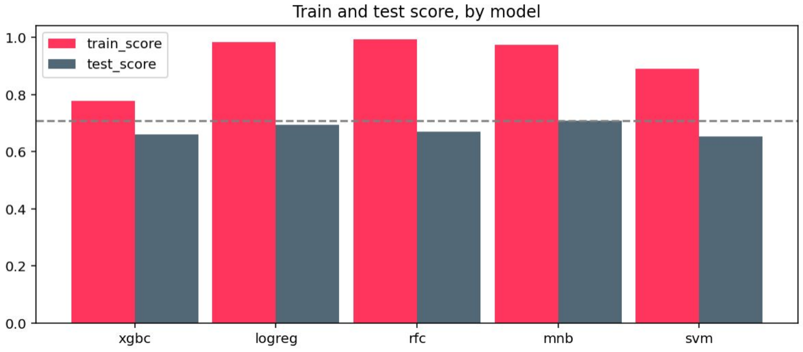


- Only 20 of the top 40 trigrams used by men and top 40 used by women were common to both Subreddits

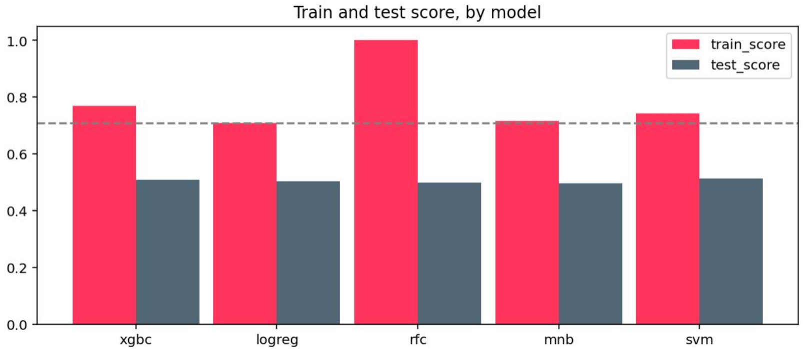
‘* Visualisation includes top 40 trigrams used in each Subreddit

# Performance of “base” models

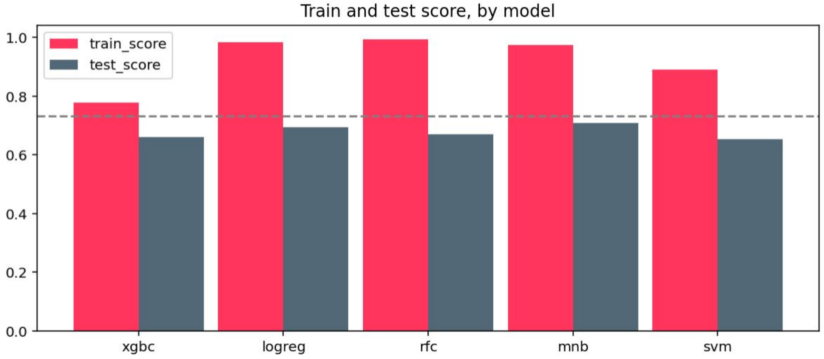
## Models using only text



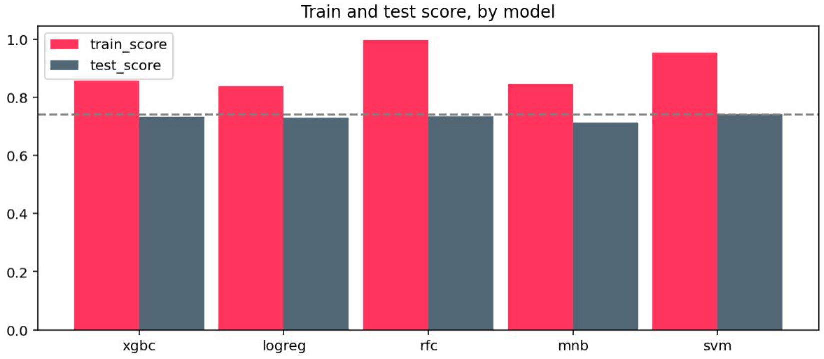
## Models using text + sentiment scores



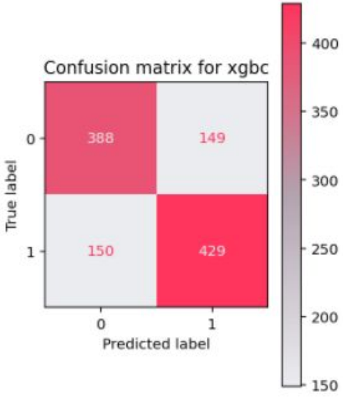
## Models using text + emoji counts

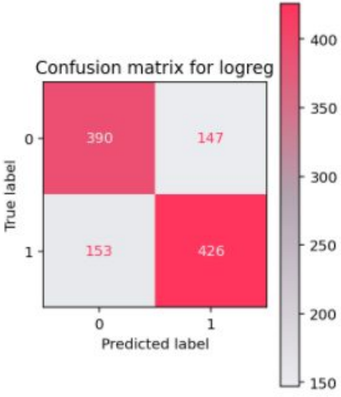


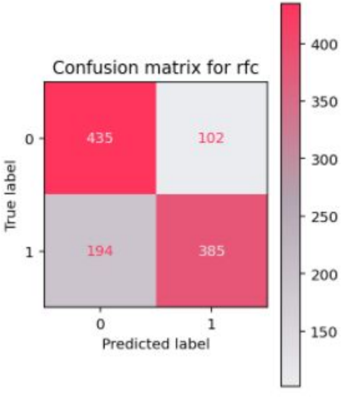
## Models using text + emoji

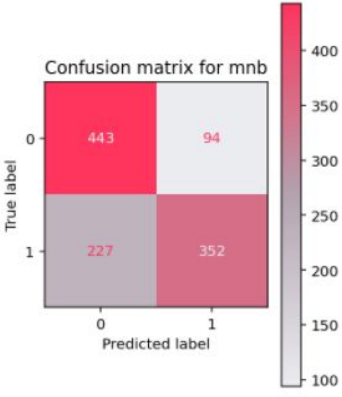


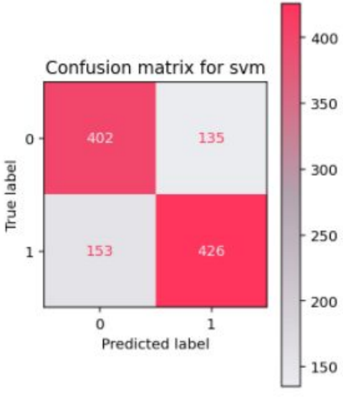
# Classification models using text + emoji









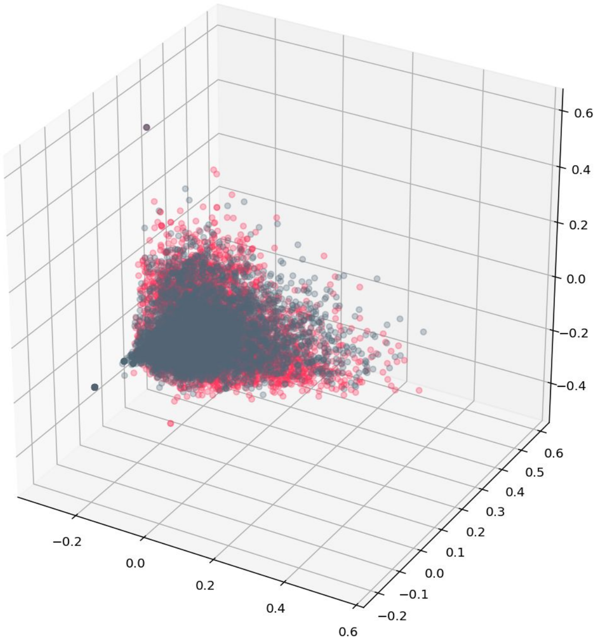


GridSearch optimization of **xgbc**, **rfc**, and **svm** hyperparameters

Voting Classifier using **xgbc** + **rfc** + **svm** with hyperparameters optimized through GridSearch

# True labels vs voting classifier predictions

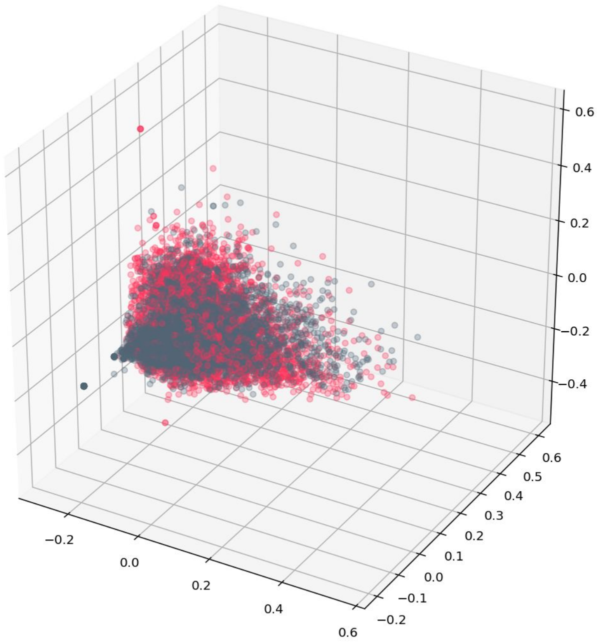
### True labels



| | | | |
|---|---|---|---|
| True label | 0 | 4,879 | 605 |
| | 1 | 708 | 4,959 |
| | | 0 | 1 |
| | Predicted label | | |

Current model accuracy = 88.2%

### Predicted labels



# Next steps

- Continue to fine-tune the hyperparameters using text and emoji as features
- Harvest more data from both subreddits or other forums

- Other suggestions?

# Q&A