# SUBREDDIT CLASSIFICATION

FAQ ON MARRIAGE ADVICE AND RELATIONSHIPS

# PROBLEM STATEMENT

- Couples who are worried about their marriage, but too reluctant to go to a marriage counselor

- Create an FAQ for married couples and soon-to-be married couples

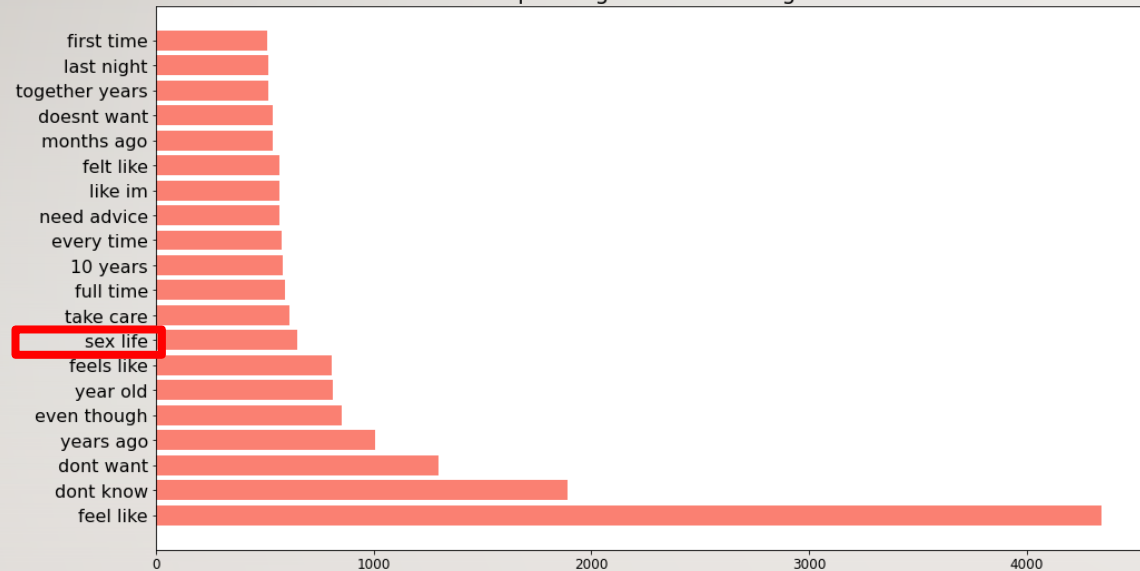- Advise them to come for a session if they are clicking into some of the more worrying FAQs.

# SUBREDDITS

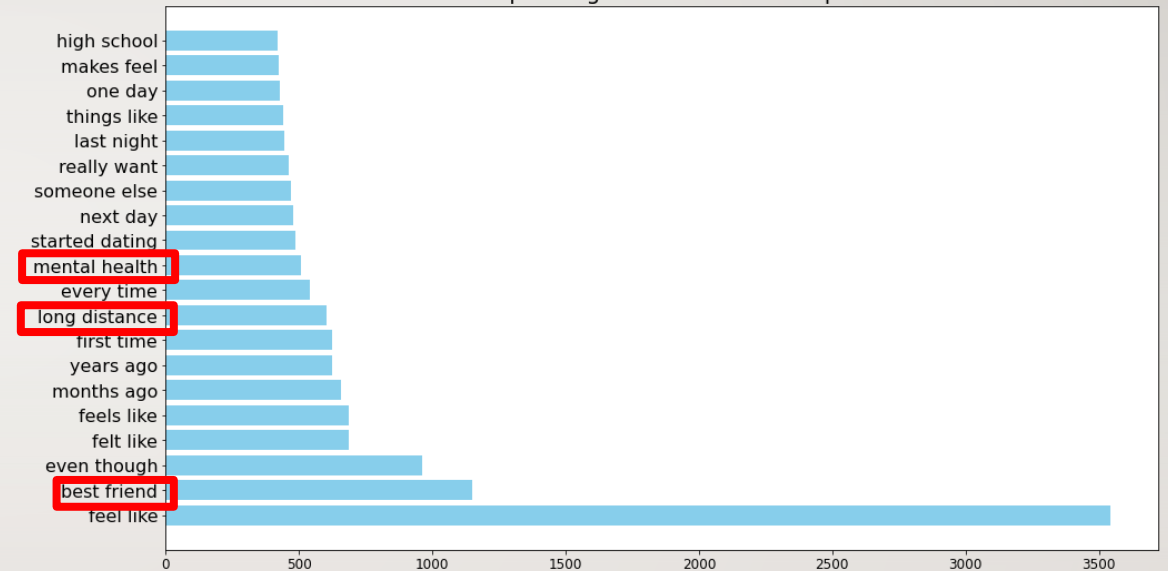Marriage Advice – Married couples seeking advice on issues they are facing.

Relationships – Other couples seeking relationship advice on issues.

# EDA – TOP FREQUENCY OF BIGRAMS
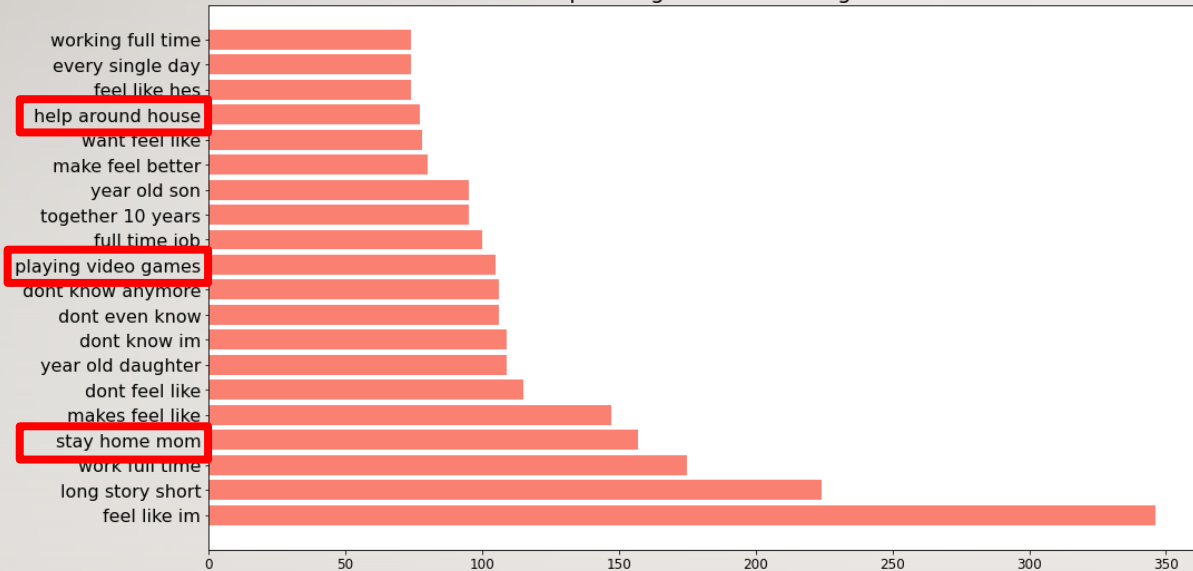


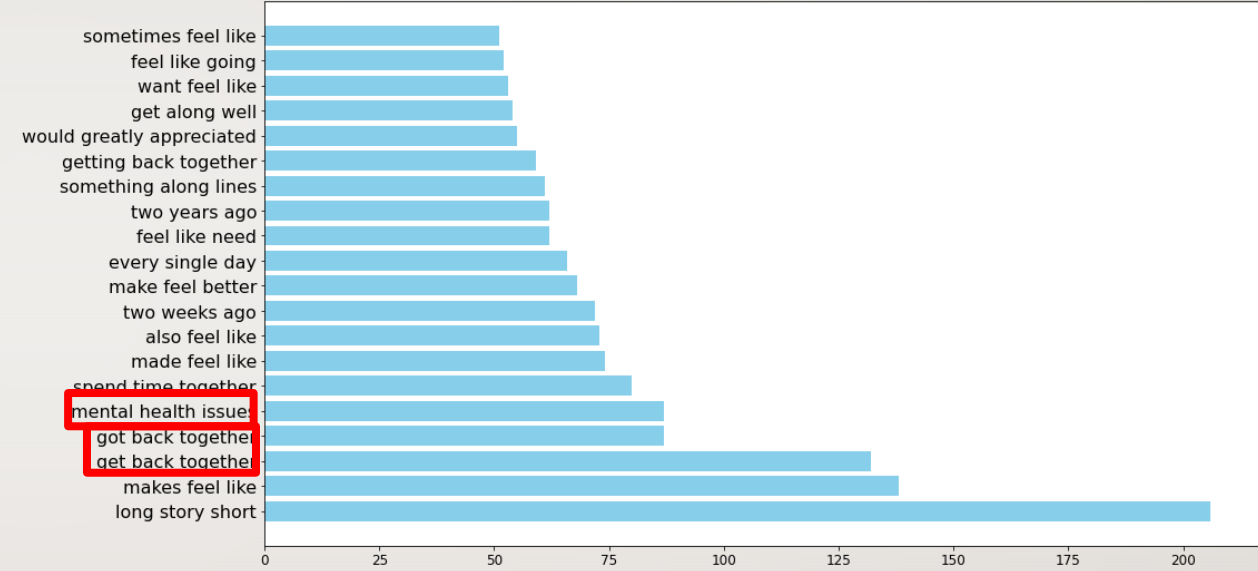Top 20 bigrams for Marriage

Top 20 bigrams for Relationship

# EDA – TOP FREQUENCY OF TRIGRAMS



Top 20 trigrams for Marriage

Top 20 trigrams for Relationships

# SUMMARY OF MODELS

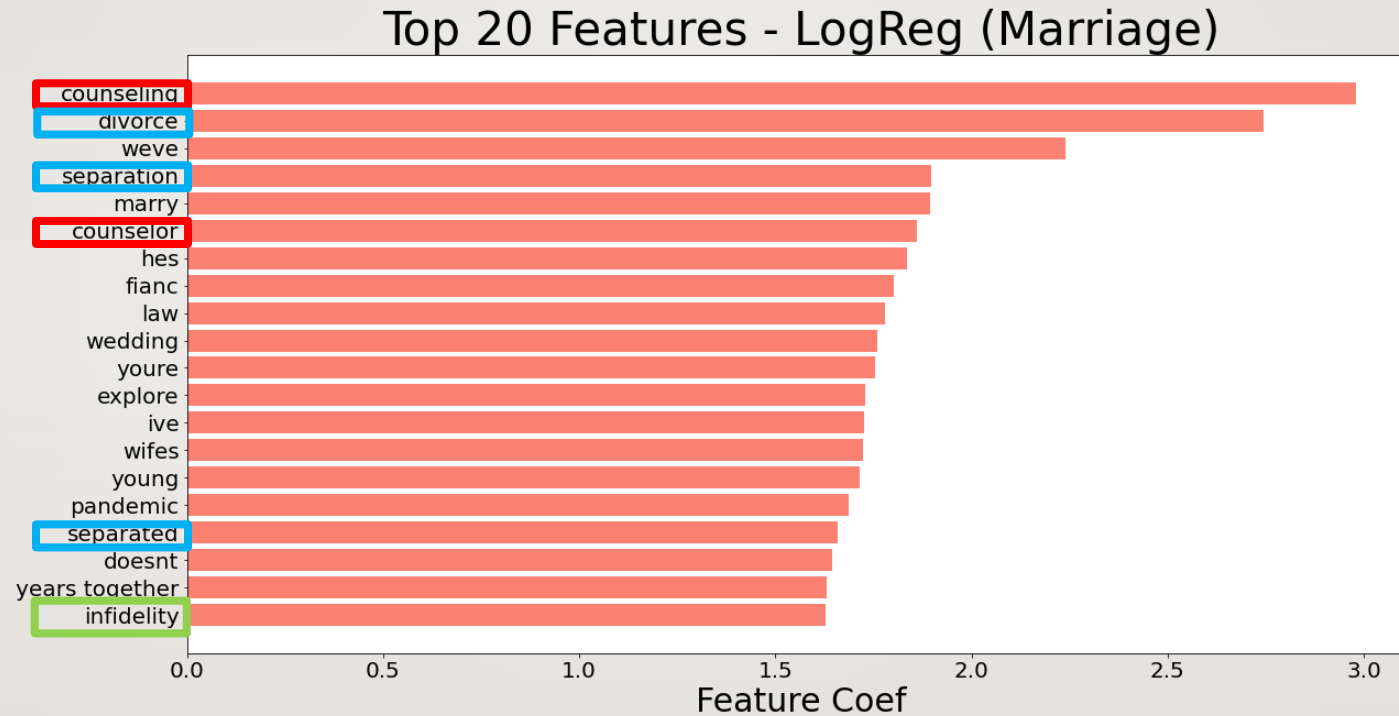| Model | Vectorizer | Train score | Test Score | ROC AUC Score | TN | FP | FN | TP |
|---|---|---|---|---|---|---|---|---|
| Benchmark Model | CVEC | 89.3% | 83.4% | 90.3% | 1446 | 322 | 280 | 1583 |
| MultinomialNB | CVEC | 86.4% | 83.6% | 90.2% | 1460 | 308 | 288 | 1575 |
| MultinomialNB | TVEC | 86.8% | 83.4% | 91.3% | 1437 | 331 | 271 | 1592 |
| LogisticRegression | CVEC | 94.7% | 84.7% | 91.8% | 1466 | 302 | 255 | 1608 |
| LogisticRegression | TVEC | 91.0% | 85.1% | 92.7% | 1477 | 291 | 252 | 1611 |
| RandomForest | CVEC | 100% | 83.4% | 91.9% | 1460 | 308 | 294 | 1569 |
| RandomForest | TVEC | 99.99% | 82.9% | 91.4% | 1453 | 315 | 305 | 1558 |

- **CountVectorizer**
  - max_df=0.95,
  - max_features=6000,
  - min_df=2,
  - ngram_range=(1, 2),
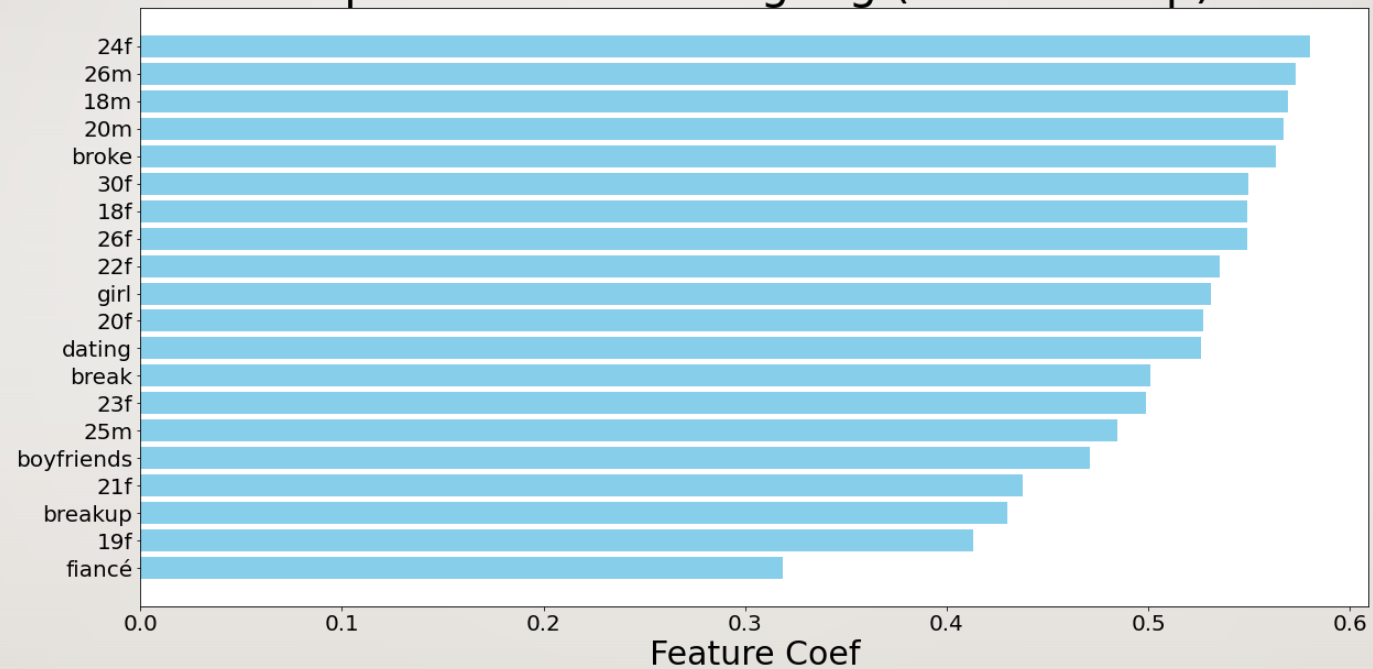  - stop_words,

- **LogisticRegression**
  - C=0.1,
  - max_iter=1000,
  - Penalty=l1
  - solver='liblinear'

# TOP FEATURES FOR MARRIAGE



Top 20 Features - LogReg (Marriage)

# TOP FEATURES FOR RELATIONSHIPS



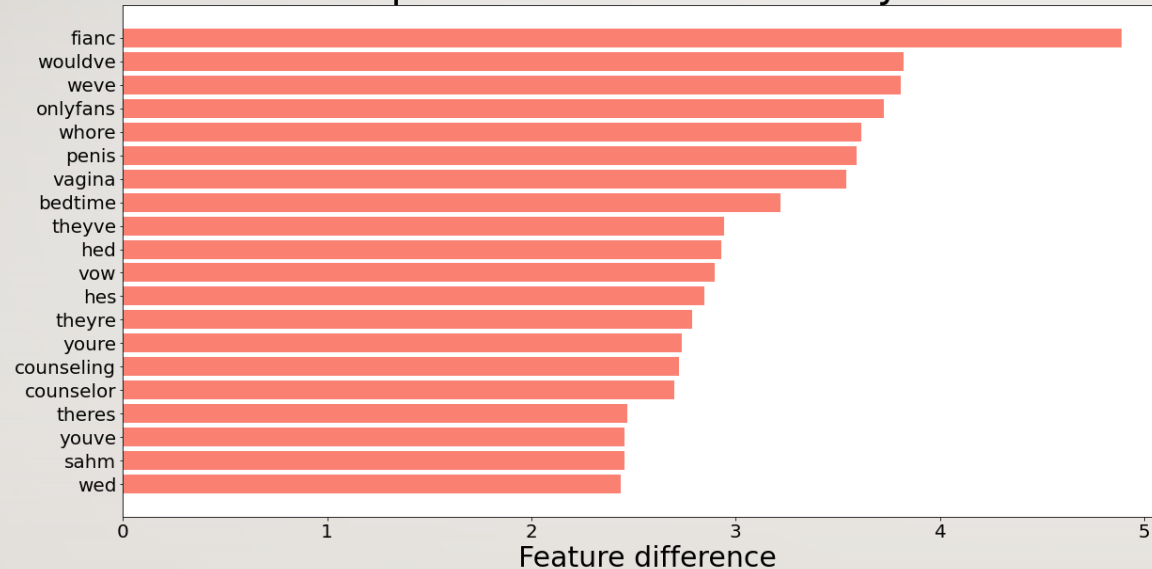Top 20 Features - LogReg (Relationship)

# ERROR ANALYSIS & TRADEOFFS

- Some married couples had posted in Relationship reddit instead.
  - Approximately 400 posts (out of 8800+)

- Using CVEC instead of TVEC, accuracy was slightly reduced.

- CVEC eventually more useful to my problem statement.

- After modeling, words identified in EDA are lost

- Having only one model limits the extend of being able to create an extensive FAQ.

# TOP FEATURES FOR NAÏVE BAYES



Top 20 Features - Naive Bayes

- Title: "Husband sneaking around online"

- Post:

"So recently we got a suspicious charge from onlyfans on our account. I could see he tried to hide it but when I confronted him, he said he just didn't know what it was or how it worked (yeah right) so then I did a little research and found out that he has been commenting on naked pictures of women saying how hot they are and asking for their Snapchat. Is this cheating? I feel like I have been cheated on and He says it means nothing."

# AREAS FOR FUTURE WORK

- Sentiment Analysis of the Posts using tailored Sentiment Library


- Classify posts by age of user

# THANK YOU