# Right Classifying Subreddit Posts from Moms and Dads through Natural Language Processing (NLP)

Prepared by: Lindy Tan

# Problem Statement

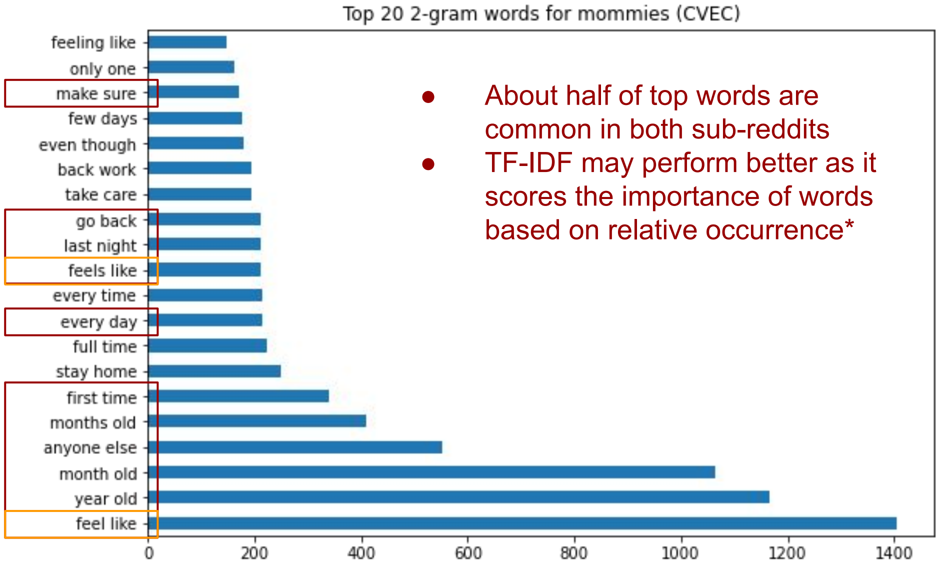
- With increasing number of working moms, and dads having to play an increasing role in taking care of children, child rearing becomes increasing a shared responsibility

- As an employee of theAsianparent, we help parents experience healthy pregnancies and raise healthy families. Through this project, we hope to:
  a. Automate the classification and monitoring of posts and discussions by moms and dads
  b. Better understand concerns and topics of interest amongst them to create targeted and interesting articles to better support parenthood

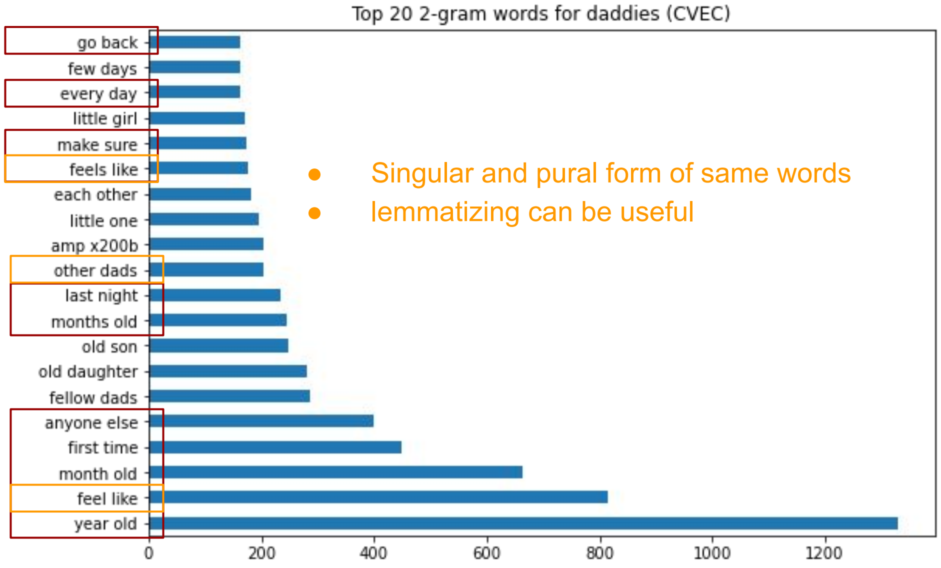
- Target Audience: Moms and dads

- We will train our model using over 10,000 posts each from r/mommit and r/daddit

# Insights from EDA: Many common words and presence of singular and plural form of same words



- About half of top words are common in both sub-reddits
- TF-IDF may perform better as it scores the importance of words based on relative occurrence*

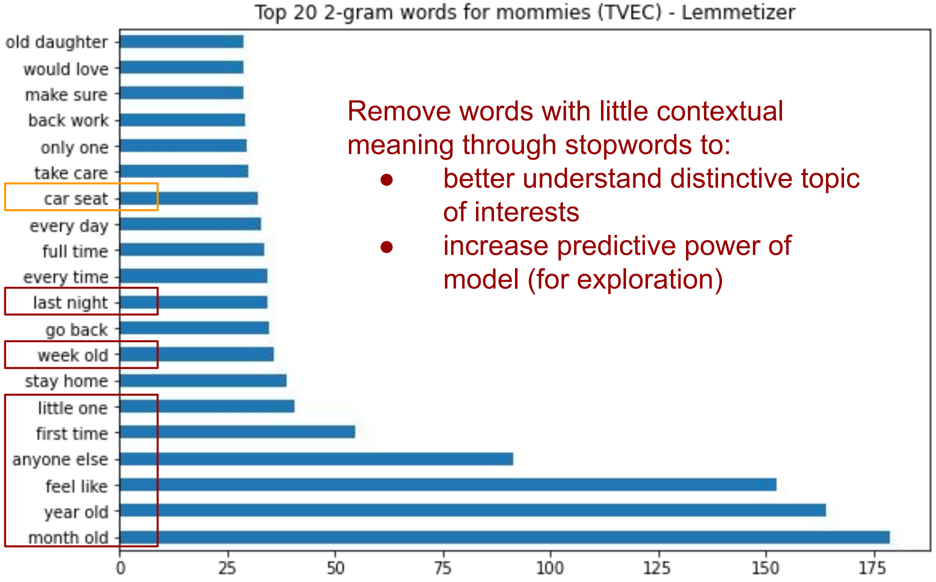


- Singular and pural form of same words
- lemmatizing can be useful

* TF-IDF may perform better as it scores the importance of words based on its relative occurrence in the doc as compared to all other docs. Words which occur often in both reddits will have lower importance while those which occur in one document but not in many documents are more important and contain more predictive power.

# Insights from EDA: While TF-IDF reduces the ‘score’ of common words, they still remain as top words which can be removed for clearer insights

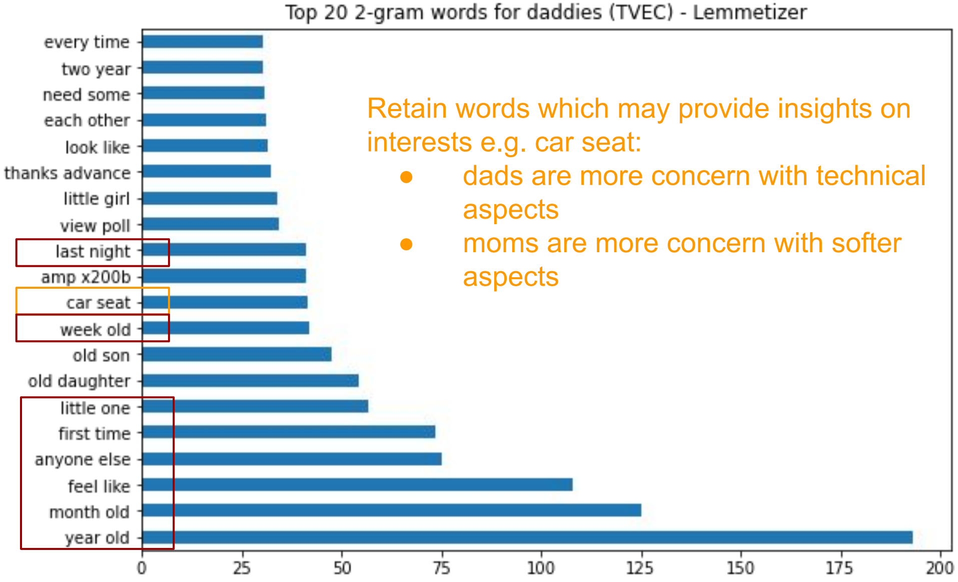
- Some of these words have little contextual meaning and hinder our ability to understand the dads' and moms' top concerns and topics of interest



Top 20 2-gram words for mommies (TVEC) - Lemmetizer

Remove words with little contextual meaning through stopwords to:

- better understand distinctive topic of interests
- increase predictive power of model (for exploration)

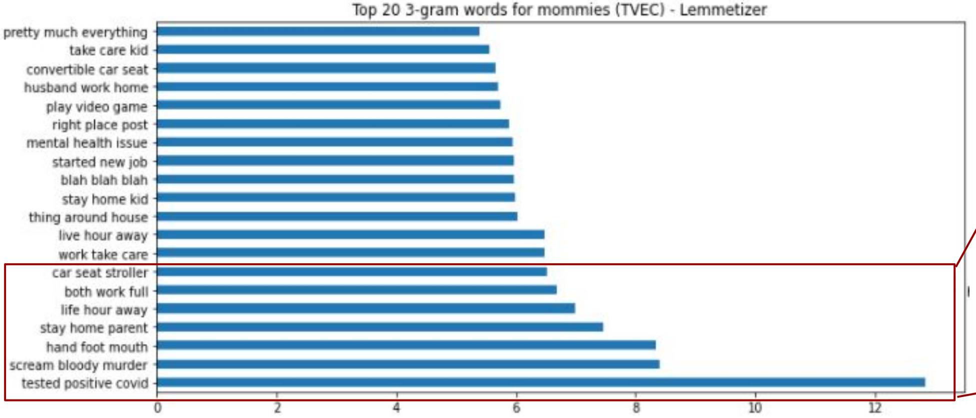


Top 20 2-gram words for daddies (TVEC) - Lemmetizer

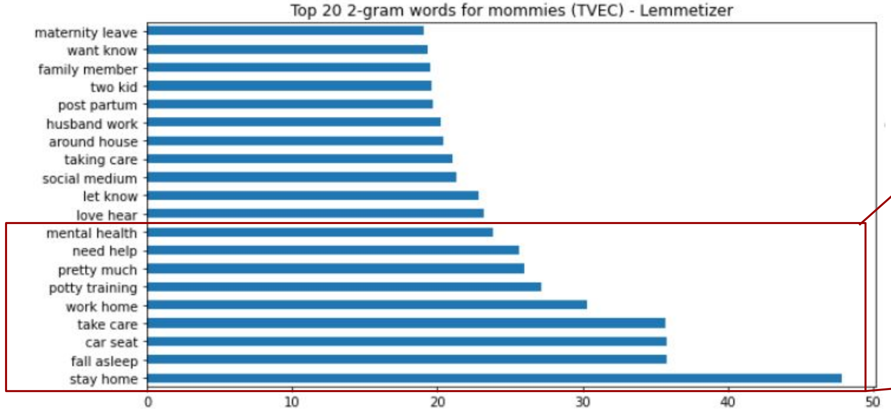
Retain words which may provide insights on interests e.g. car seat:

- dads are more concern with technical aspects
- moms are more concern with softer aspects

# Insights from EDA: Moms' concerns and topics of interest are on COVID, car seat, potty training and postpartum mental health. Also talk about being stay-home parents.

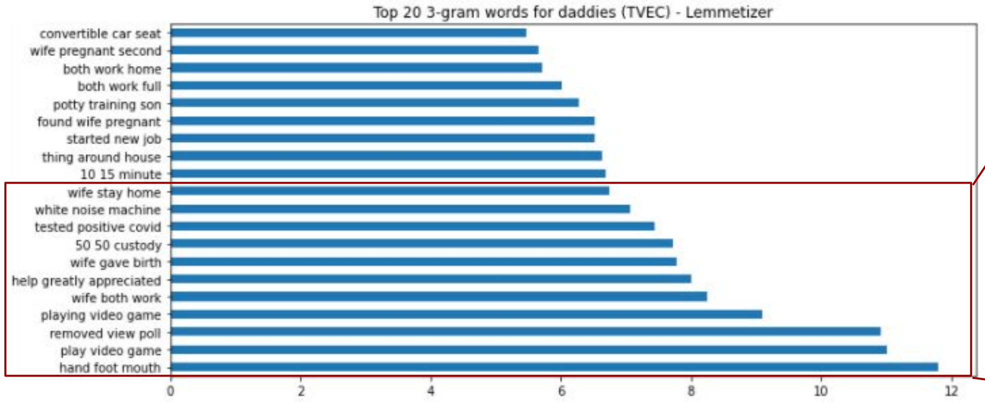


- Baby or themself **tested COVID positive**
- Taming crying baby (scream bloody murder)
- Hand foot and mouth disease (HFMD)
- **Car seat and stroller** travel system
- Describing themselves as **stay-home** or working full-time moms/parents

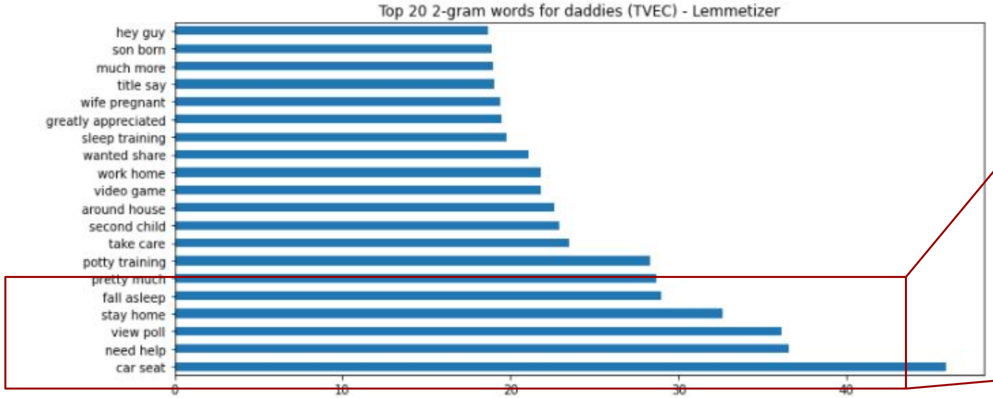


- Making baby fall asleep (on their own)
- **Potty-training** their child
- **Postpartum mental health**

# Insights from EDA: Dads’ concerns and discussion topics are on video games, COVID, potty training and wife’s pregnancy. Also talk about them/ wives being stay-home parent.



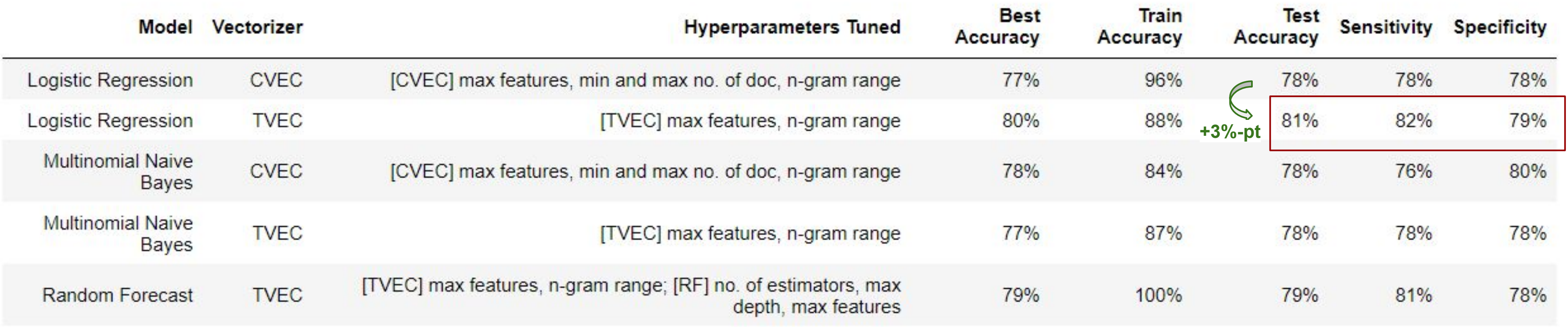
- HFMD
- **Playing video games** with their children
- Wife **pregnant or giving birth**
- 50/50 custody for divorcees
- **Tested COVID positive**
- White noise machine to help baby sleep
- ***Like to conduct polls***
- Describing their wives and themselves as **stay-home** or working full-time parents



- Showing how their child fall asleep (and how they fall asleep with them)
- Seeking advice and sharing success stories of **potty training**

# Model Exploration: Logistic Regression and Multinomial Naive Bayes with count vectorizer and TF-IDF vectorizer, as well as Random Forecast on TD-IDF vectorizer

- Based on Logistic Regression and Multinomial Naive Bayes, TF-IDF vectorizer is comparable or, if not, outperform Count Vectorizer
  - ○ This is due to presence of many common terms in both subreddits
  - ○ Extended the use of TF-IDF for Random Forest



| Model | Vectorizer | Hyperparameters Tuned | Best Accuracy | Train Accuracy | Test Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|
| Logistic Regression | CVEC | [CVEC] max features, min and max no. of doc, n-gram range | 77% | 96% | 78% | 78% | 78% |
| Logistic Regression | TVEC | [TVEC] max features, n-gram range | 80% | 88% | 81% | 82% | 79% |
| Multinomial Naive Bayes | CVEC | [CVEC] max features, min and max no. of doc, n-gram range | 78% | 84% | 78% | 76% | 80% |
| Multinomial Naive Bayes | TVEC | [TVEC] max features, n-gram range | 77% | 87% | 78% | 78% | 78% |
| Random Forecast | TVEC | [TVEC] max features, n-gram range; [RF] no. of estimators, max depth, max features | 79% | 100% | 79% | 81% | 78% |

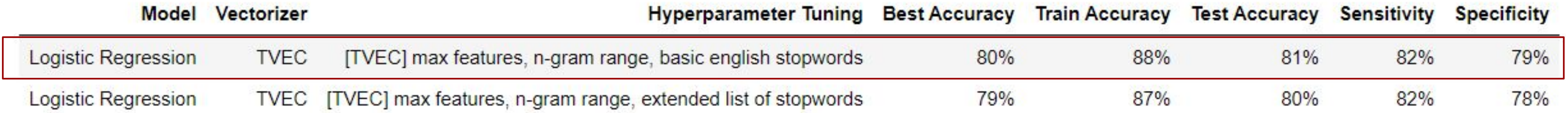
+3%-pt

Logistic Regression with TF-IDF vectorizer is the best model as it has the highest accuracy, sensitivity and specificity

- ○ Accuracy increased by 3%-pt compared to base model
- ○ Although specificity is marginally lower than Multinomial Naive Bayes with countvectorizer, both accuracy and sensitivity are higher

# Model Exploration: Logistic Regression with TF-IDF vectorizer on basic english stopwords (including obvious and associated words) and extended list of stopwords

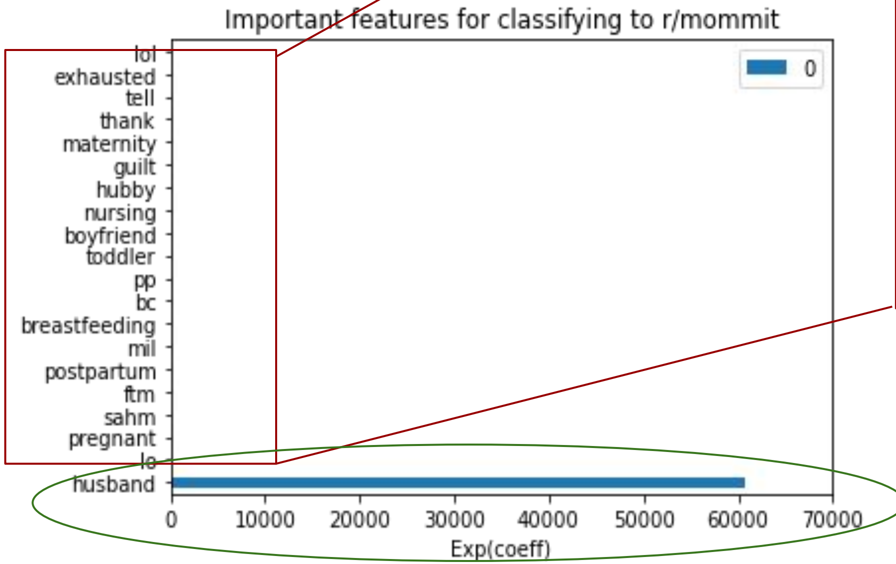
- Explored further tuning the best model on extended list of stopwords to remove words which are common but have little contextual meaning
- Extended list of stopwords does not help to increase accuracy, sensitivity and specificity of model
  - Likely due to the removal of words with low feature importance (i.e. high occurrence across both subreddits)

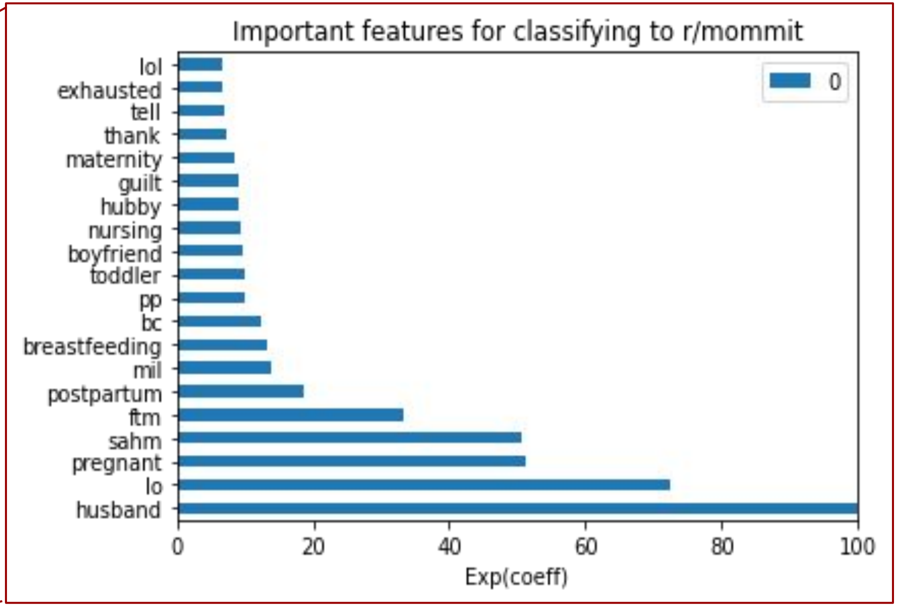


| Model | Vectorizer | Hyperparameter Tuning | Best Accuracy | Train Accuracy | Test Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|
| Logistic Regression | TVEC | [TVEC] max features, n-gram range, basic english stopwords | 80% | 88% | 81% | 82% | 79% |
| Logistic Regression | TVEC | [TVEC] max features, n-gram range, extended list of stopwords | 79% | 87% | 80% | 82% | 78% |

Final Model: Logistic Regression with TF-IDF vectorizer on basic english stopwords with max features of 9,000 and n-gram range of (1, 2)

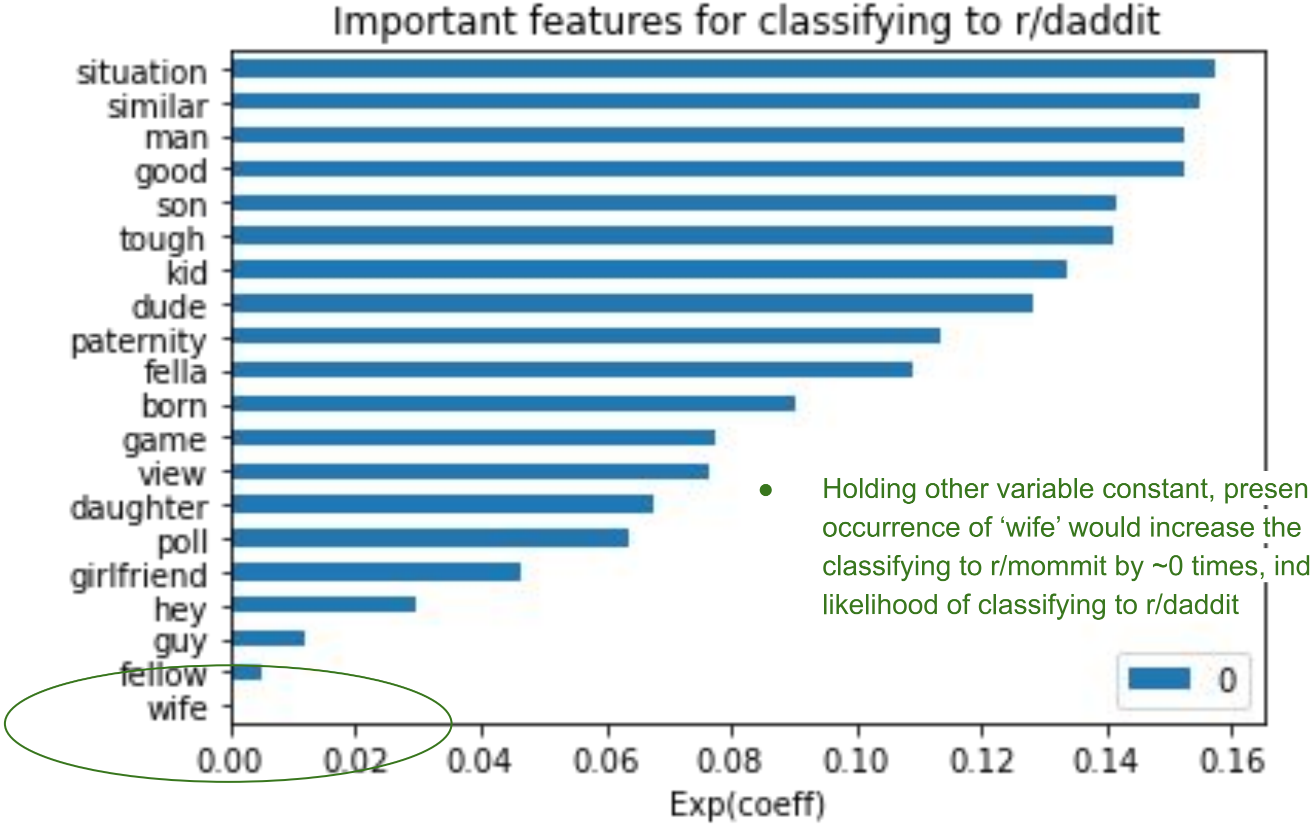
# Important features for classifying posts to r/mommit are husband, pregnant, SAHM, FTM, postpartum and breastfeeding. These are more distinctive words used by moms.





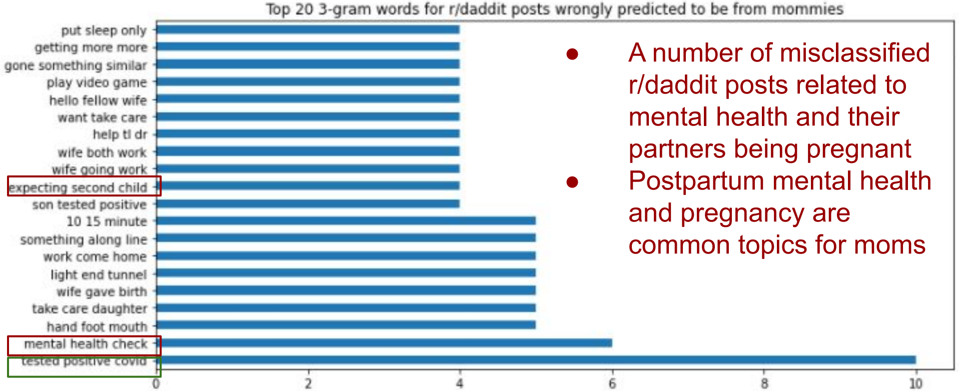
- Logistic regression allows us to interpret the importance of each feature through the coefficient size
  - For example, holding other variable constant, presence of one occurrence of ‘husband’ would increase the likelihood of classifying to r/mommit by ~60,000 times.

For r/daddit, the important features and distinctive words used include wife, fellow, guy, hey, girlfriend and game.

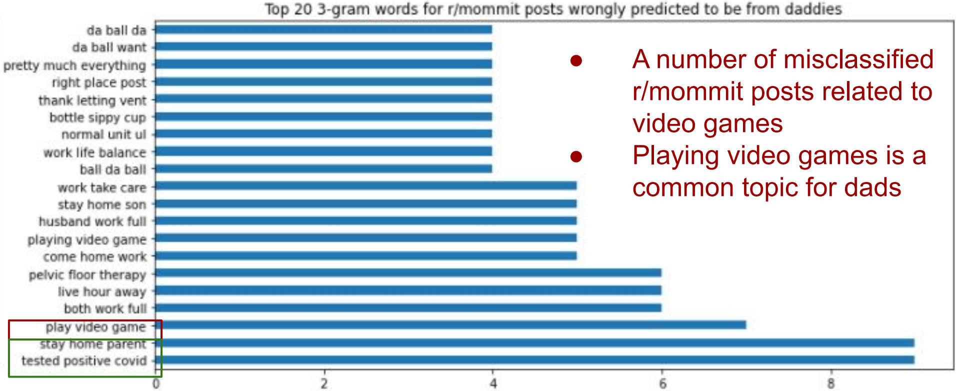


- Holding other variable constant, presence of one occurrence of ‘wife’ would increase the likelihood of classifying to r/mommit by ~0 times, indicating high likelihood of classifying to r/daddit

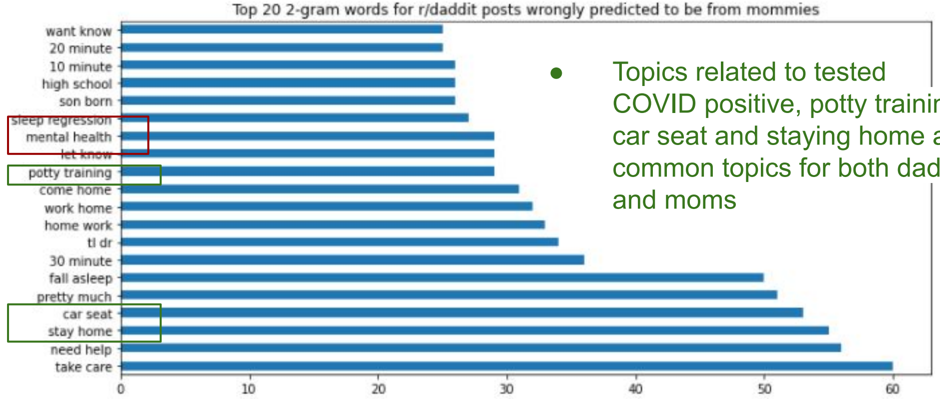
# Error Analysis: Based on top words of wrongly classified posts, many contain important features (more distinctive discussion topics) from the other subreddit



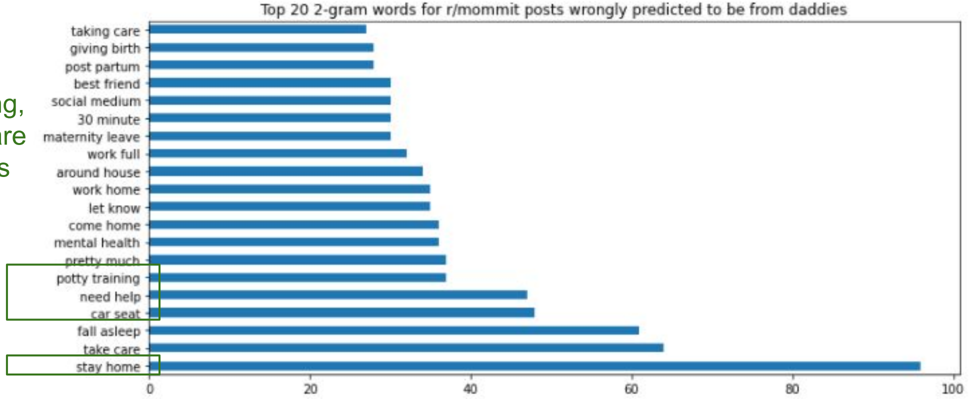
- A number of misclassified r/daddit posts related to mental health and their partners being pregnant
- Postpartum mental health and pregnancy are common topics for moms



- A number of misclassified r/mommit posts related to video games
- Playing video games is a common topic for dads



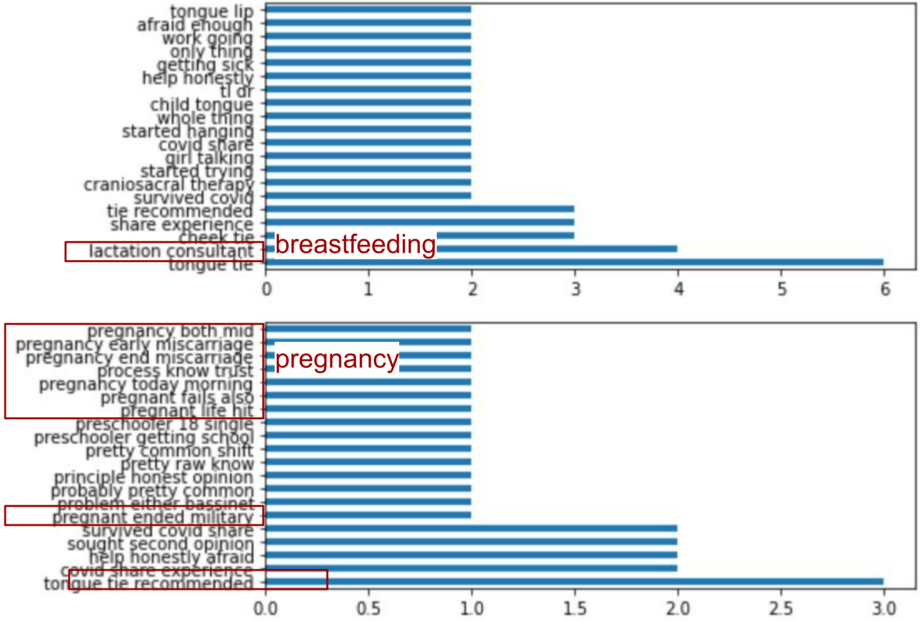
- Topics related to tested COVID positive, potty training, car seat and staying home are common topics for both dads and moms



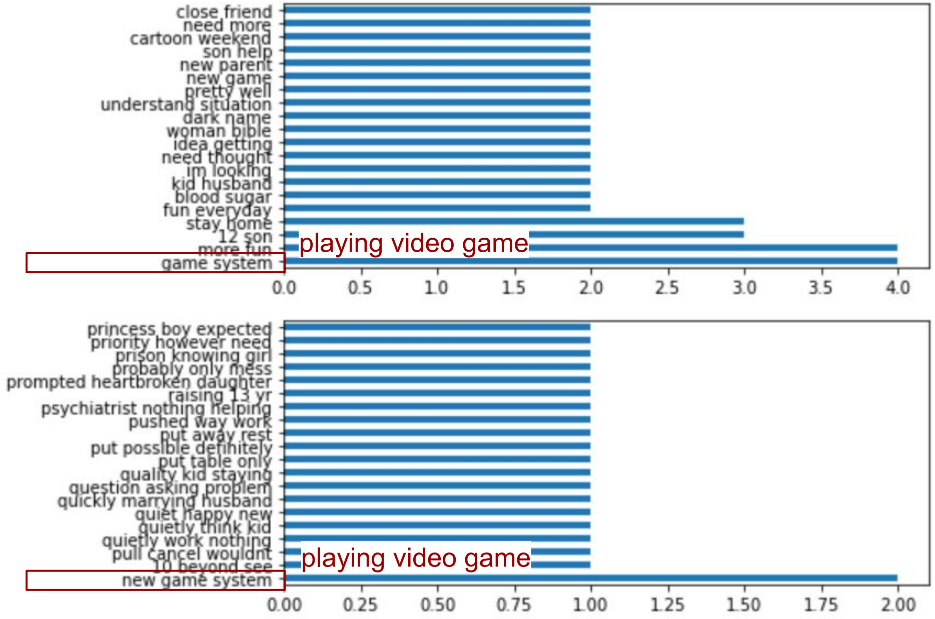
# Error Analysis: Despite containing term ‘husband’ and ‘wife’, ~5% of posts from each subreddit are wrongly classified because of long length with topics of other subreddit

- Based on strength of feature importance, it is unlikely for posts containing ‘husband’ or ‘wife’ to be wrongly classified
- However, ~5% of posts in each subreddits are misclassified despite mentioning  ‘husband’ or ‘wife’
  - These posts are usually long and contain discussion topics of the another subreddit

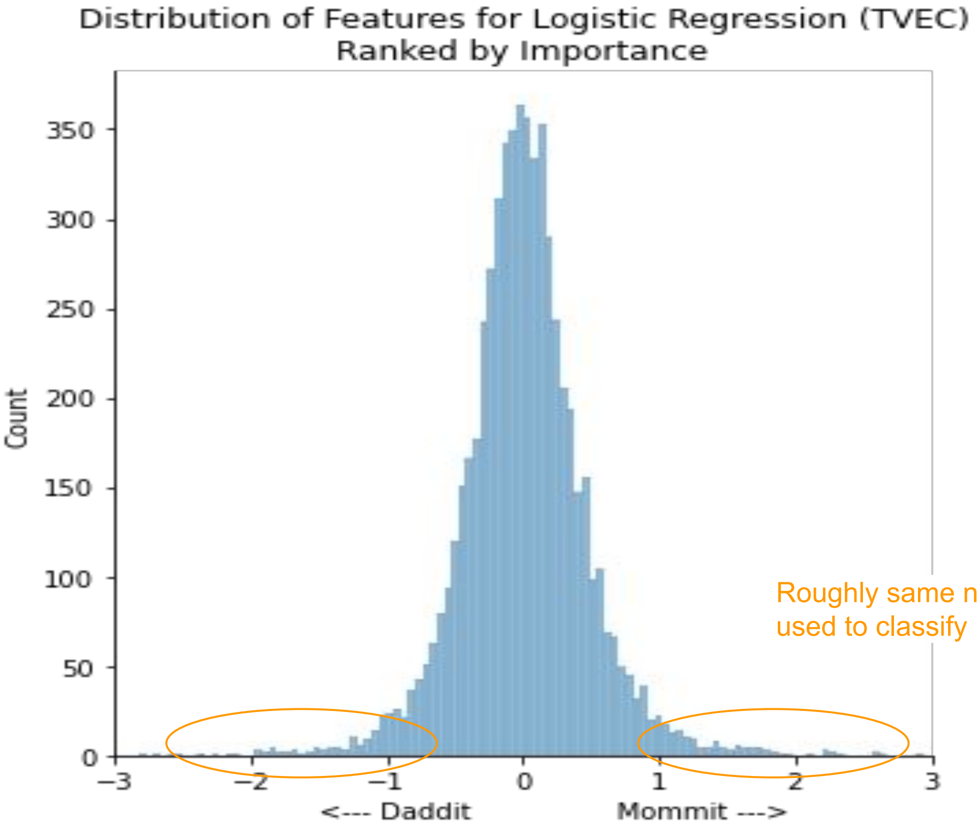
Top words for wrongly classified r/daddit posts despite mentioning of "wife"



Top words for wrongly classified r/mommit posts despite mentioning of "husband"



# Error Analysis: Feature distribution by importance is quite centralised around 0, indicating many features lack predictive power to right-classify the posts



Roughly same number of of important predictors used to classify posts to each subreddit

# Conclusion and Recommendations

- Based on error analysis, posts are wrongly classified because they are typically long and contain words relating to discussion topics from the other sub-reddit (e.g. r/daddit posts talking about breastfeeding)
- Despite this, the model performs quite well with ~80% accuracy, sensitivity (proportion of /mommit posts being right classified) and specificity (proportion of /daddit posts being right classified)
    - Allow us to reasonably right classify 80% of posts for monitoring of discussions by moms and dads
- Possible targeted topics to focus in articles for moms and dads:

| | **Moms** | **Dads** | **Both Parents** |
|---|---|---|---|
| Topics | ● Pregnancy<br>● Postpartum mental health<br>● Breastfeeding<br>● Getting husband’s help<br>● Taming crying baby<br>● Putting baby to sleep and sleep training<br>● Car seat and stroller travel system | ● Video games with children<br>● Helping wives who are pregnant or giving birth<br>● White noise machine to help baby sleep<br>● Custody for divorcees | ● Managing disease such as COVID and HFMD<br>● Potty training<br>● Stay home with children |
| Ways to engage | Providing advice | Including poll or other fun element | |

# Next Steps

- Further automate the curation of in-depth discussion areas for each topic of interests or concern for moms and dads separately (e.g. analysing posts classified as r/mommit and containing discussion on pregnancy to create content/articles relating to pregnancy).

- Explore deep-learning models to improve accurate, sensitivity and specificity as the discussions in the posts are more english- and contextual-based.