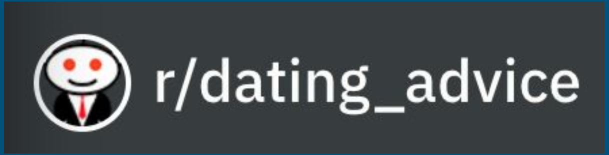

r/dating_advice

Project 3

Web APIs & Classification

Jun Yu, Jing Chun, Samuel and Vincent

RELATIONSHIP ADVICE

Executive Summary

MSF
MINISTRY OF
SOCIAL AND FAMILY
DEVELOPMENT

The social fabric of Singapore is unraveling.

Marriages are in a decline and social issues like loneliness and depression have increased exponentially. The Total Fertility Rate has plummeted to 1.0.

The Ministry of Social and Family Development has commissioned us to study the issues concerning relationships and dating.

Main Subcategories:

1) Depression and Suicide 2) Divorce 3) Dating, Marriage and Childbearing

Executive Summary

- The Ministry hopes to understand **commonplace issues** to improve therapy and counseling frameworks and outcomes.
- Leverage on **Natural Language Processing** in the future to process SOS hotlines, social media forums and the health section in the daily 'Life' publication of the Straits Times.

THE STRAITS TIMES MONDAY, OCTOBER 10, 2016

life

D

ISSA RAE
A fresh new voice on television
D3

NATHAN HARTONO
Why he is a winner in Sing! China finals
D10

Executive Summary

To get there, it needs to be able to separate **relationship** and **dating** questions and classify them in a accurate and automatic manner.

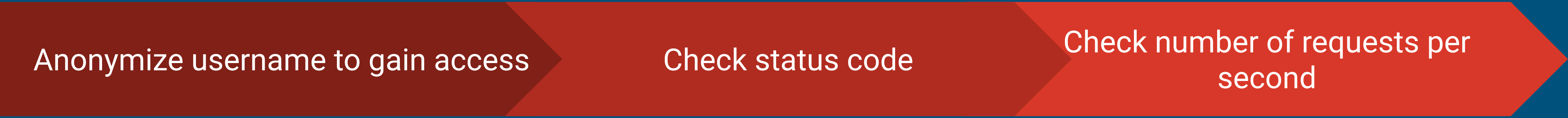
Overview Objective

- To build a web scrapper to scrape **2 subreddits (r/relationshipadivce and r/datingadvice)** to **train 2 separate classification models (CountVectorizer/TfidVectorizer and Logistic Regression)** to **predict if a post came from the respective correct subreddit.**
- Success is evaluated by ensuring that the model has the **highest accuracy scores that are consistent throughout cross validation and train/test split**

Flow of Presentation

- Data Collection
- Data Cleaning & Preprocessing
- EDA & Modelling
- Model Evaluation
- Model Limitations + Further Improvements/Exploration
- Conclusion

Data Collection

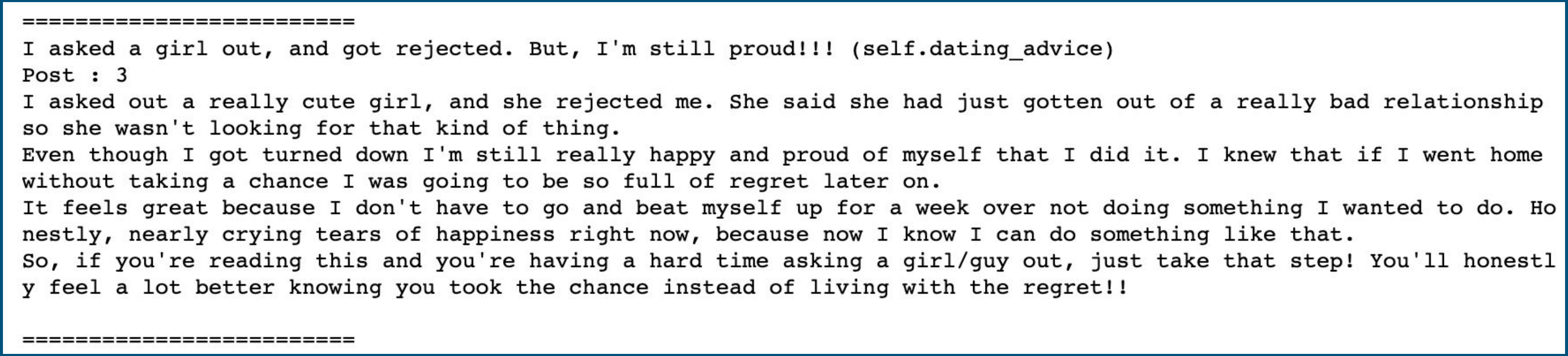


Anonymize username to gain access

Check status code

Check number of requests per second

Sample of Collection:



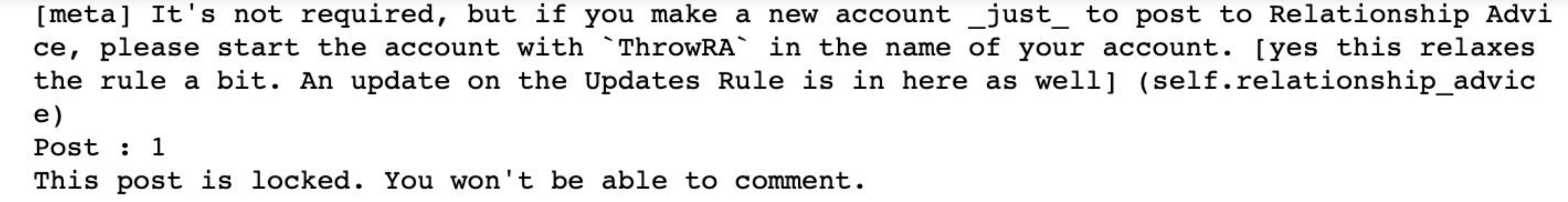
```
=========================
I asked a girl out, and got rejected. But, I'm still proud!!! (self.dating_advice)
Post : 3
I asked out a really cute girl, and she rejected me. She said she had just gotten out of a really bad relationship
so she wasn't looking for that kind of thing.
Even though I got turned down I'm still really happy and proud of myself that I did it. I knew that if I went home
without taking a chance I was going to be so full of regret later on.
It feels great because I don't have to go and beat myself up for a week over not doing something I wanted to do. Ho
nestly, nearly crying tears of happiness right now, because now I know I can do something like that.
So, if you're reading this and you're having a hard time asking a girl/guy out, just take that step! You'll honestl
y feel a lot better knowing you took the chance instead of living with the regret!!

=========================
```

Data Cleaning

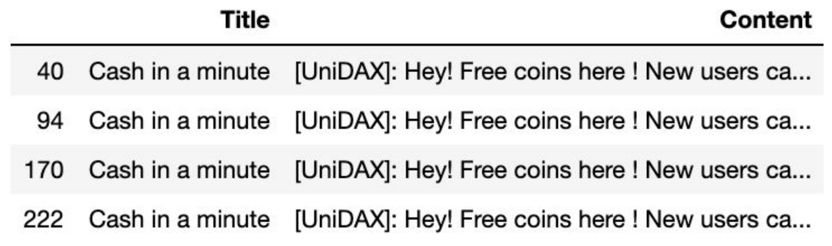
Deletion of moderator posts

Deletion of advertisements

Samples of moderator posts and advertisements::

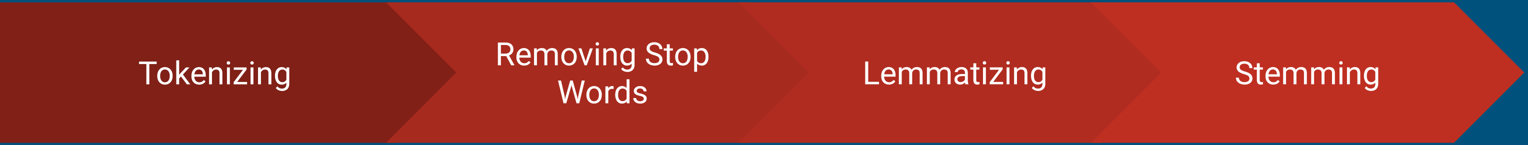


```
[meta] It's not required, but if you make a new account _just_ to post to Relationship Advi
ce, please start the account with `ThrowRA` in the name of your account. [yes this relaxes
the rule a bit. An update on the Updates Rule is in here as well] (self.relationship_advic
e)
Post : 1
This post is locked. You won't be able to comment.
```



	Title	Content
40	Cash in a minute	[UniDAX]: Hey! Free coins here ! New users ca...
94	Cash in a minute	[UniDAX]: Hey! Free coins here ! New users ca...
170	Cash in a minute	[UniDAX]: Hey! Free coins here ! New users ca...
222	Cash in a minute	[UniDAX]: Hey! Free coins here ! New users ca...

Data Preprocessing



Tokenizing

Removing Stop Words

Lemmatizing

Stemming

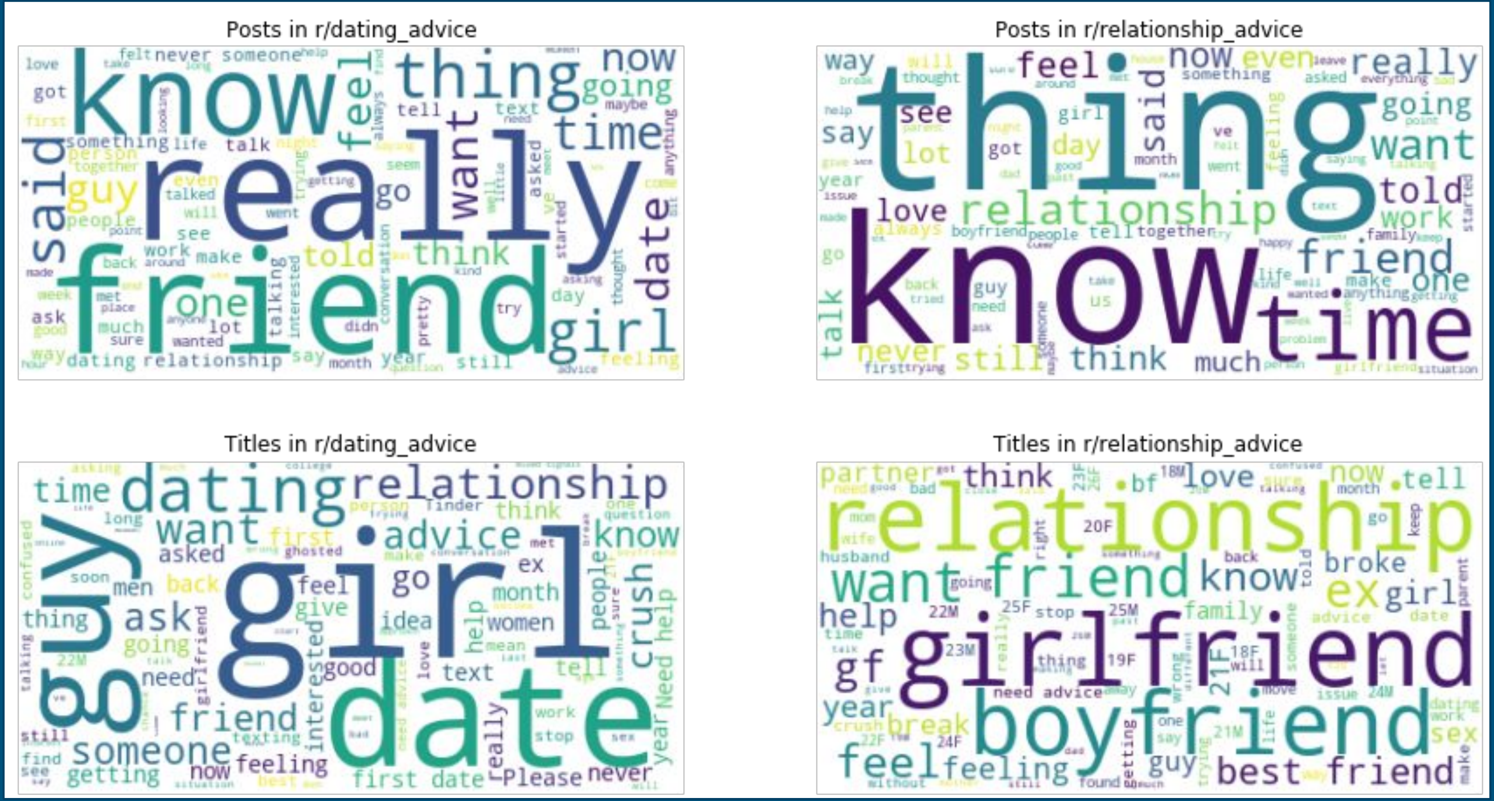
Clean Words

- Splits text into words
- Converts strings to lowercase
- Removes Punctuation

- Filters out useless words that adds little value to our data analysis
- Also filters out weird words that got corrupted in the encoding process

- Integration of words back to their common root words
- Essential in the aggregation and classification process when fitting into our CountVectorizer/TfidVectorizer

Exploratory Data Analysis



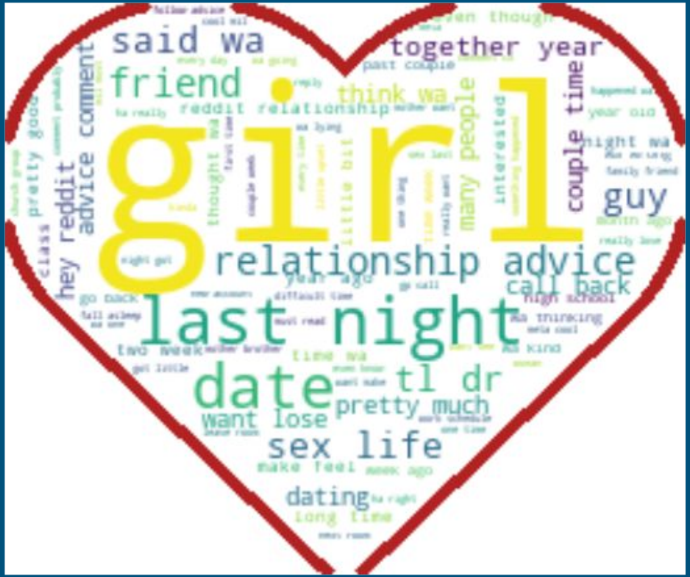
Posts in r/dating_advice

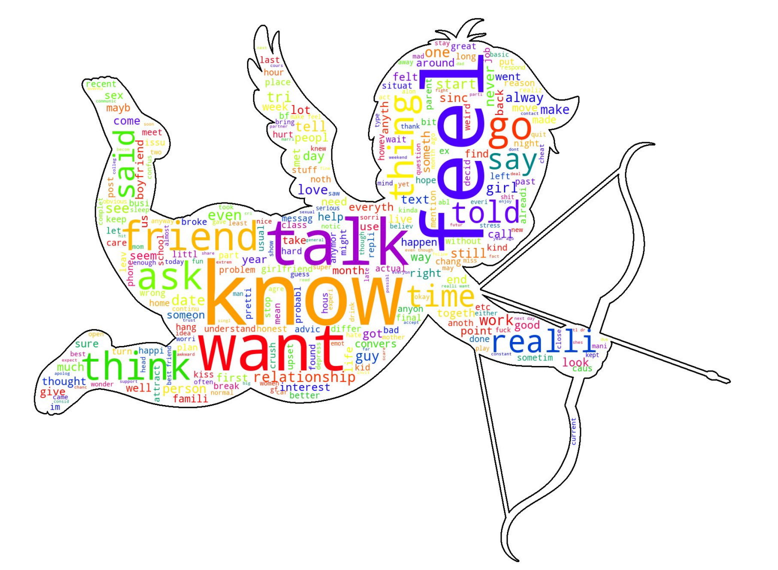
Posts in r/relationship_advice

Titles in r/dating_advice

Titles in r/relationship_advice

Exploratory Data Analysis

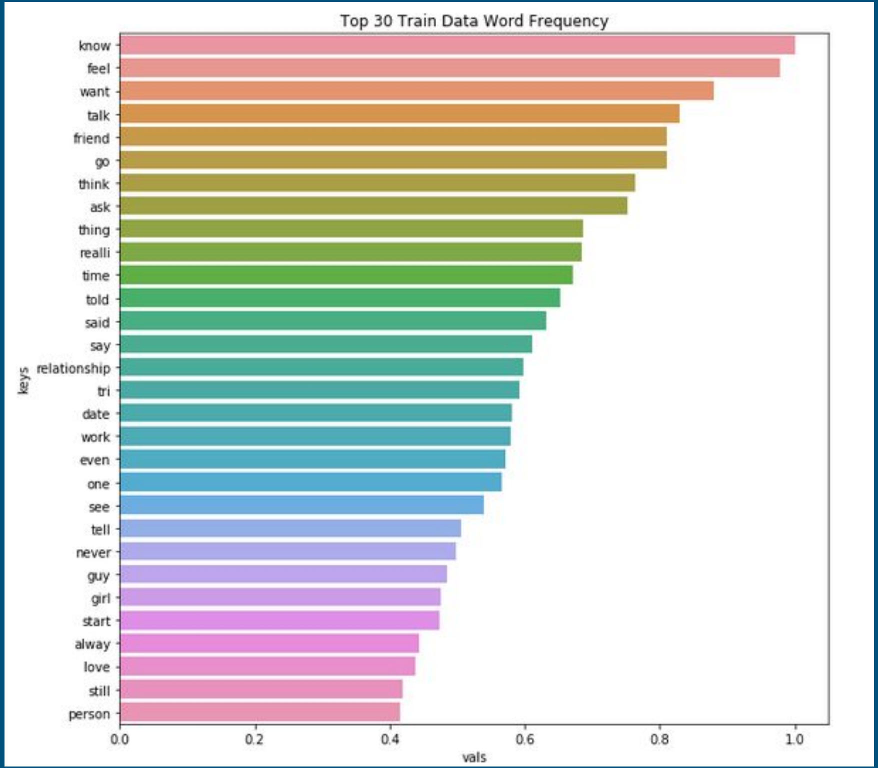


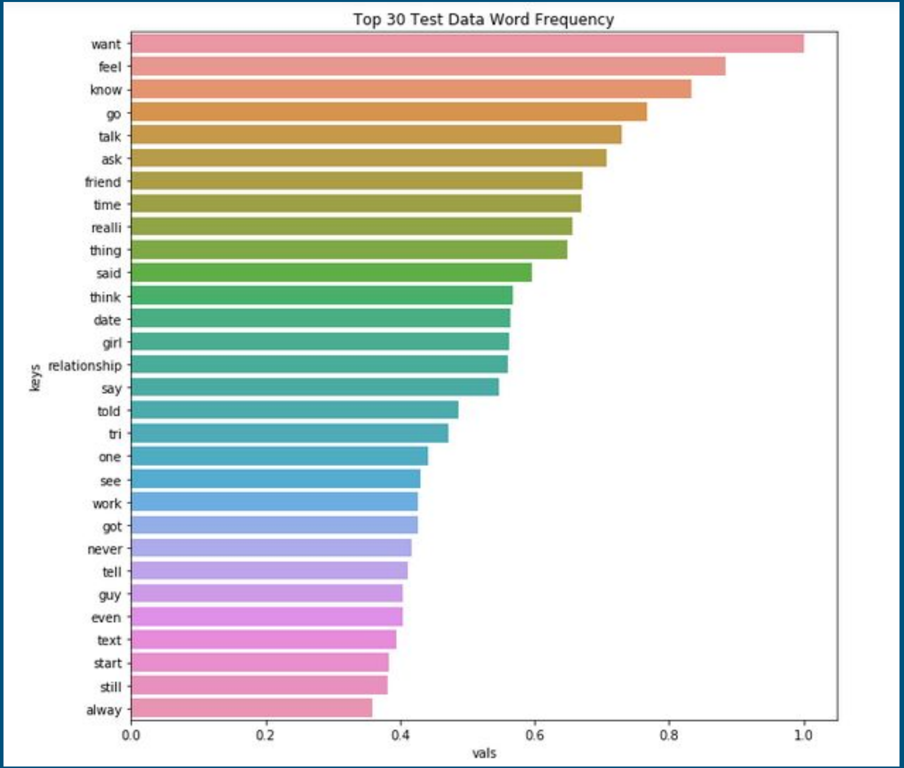


Word Cloud from concatenating title and content

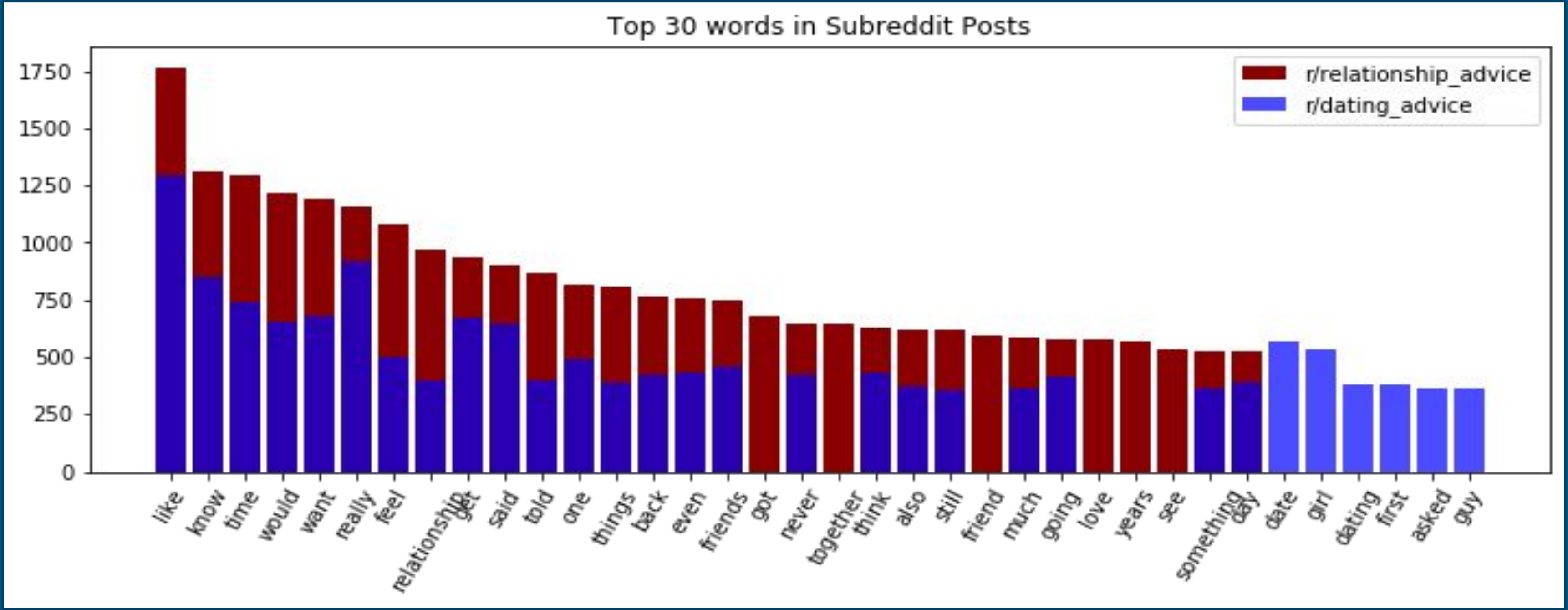
- Almost similar

Exploratory Data Analysis

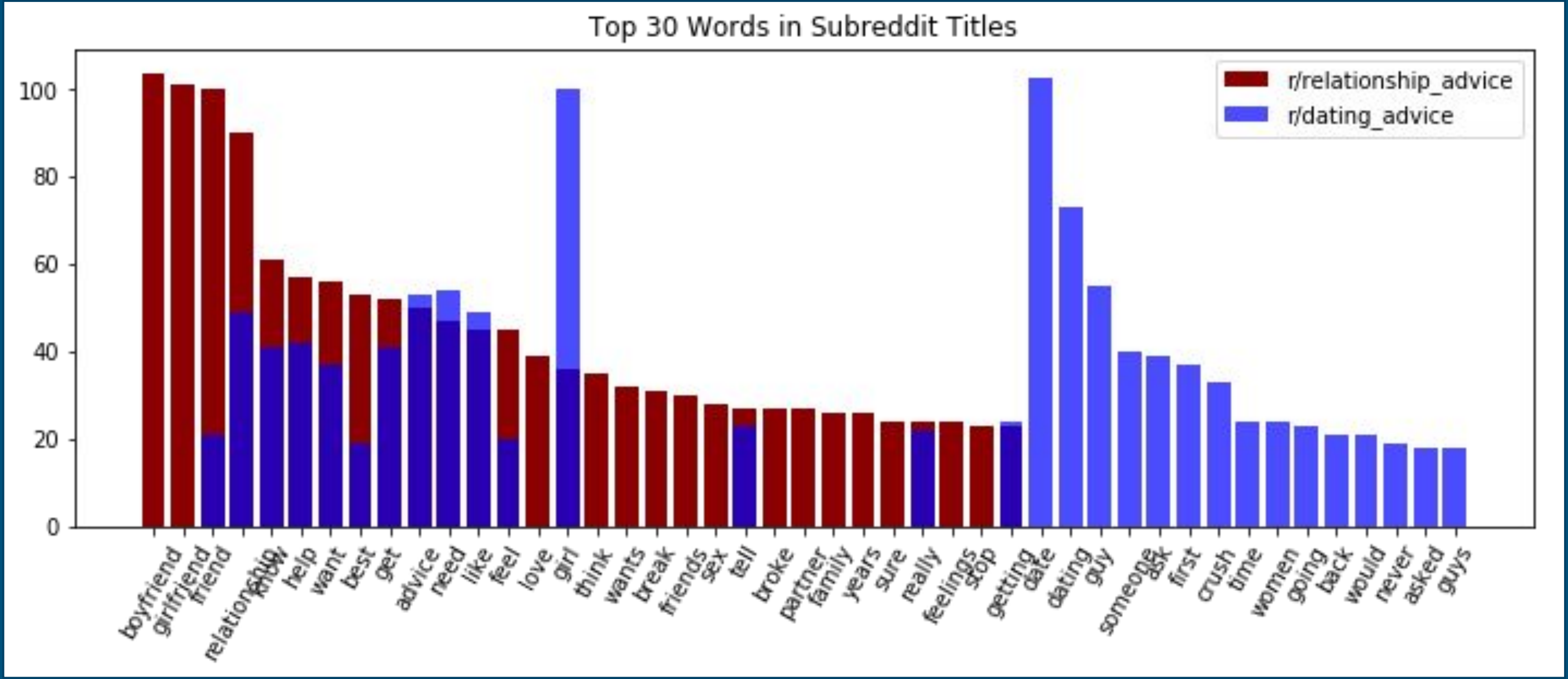




Exploratory Data Analysis

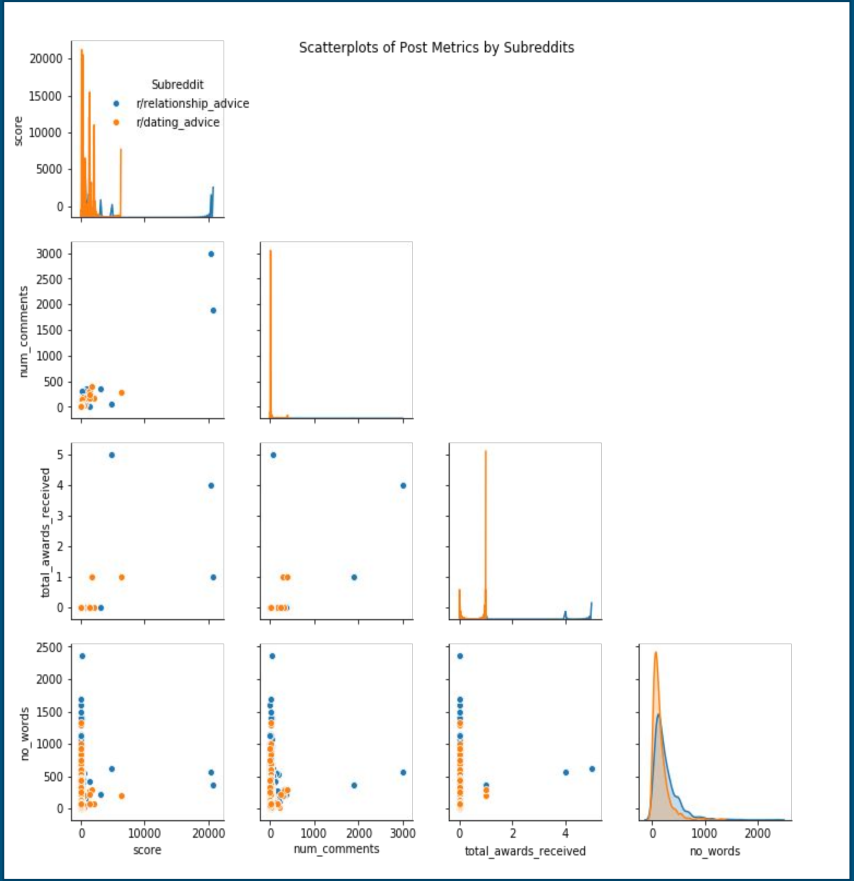


Exploratory Data Analysis

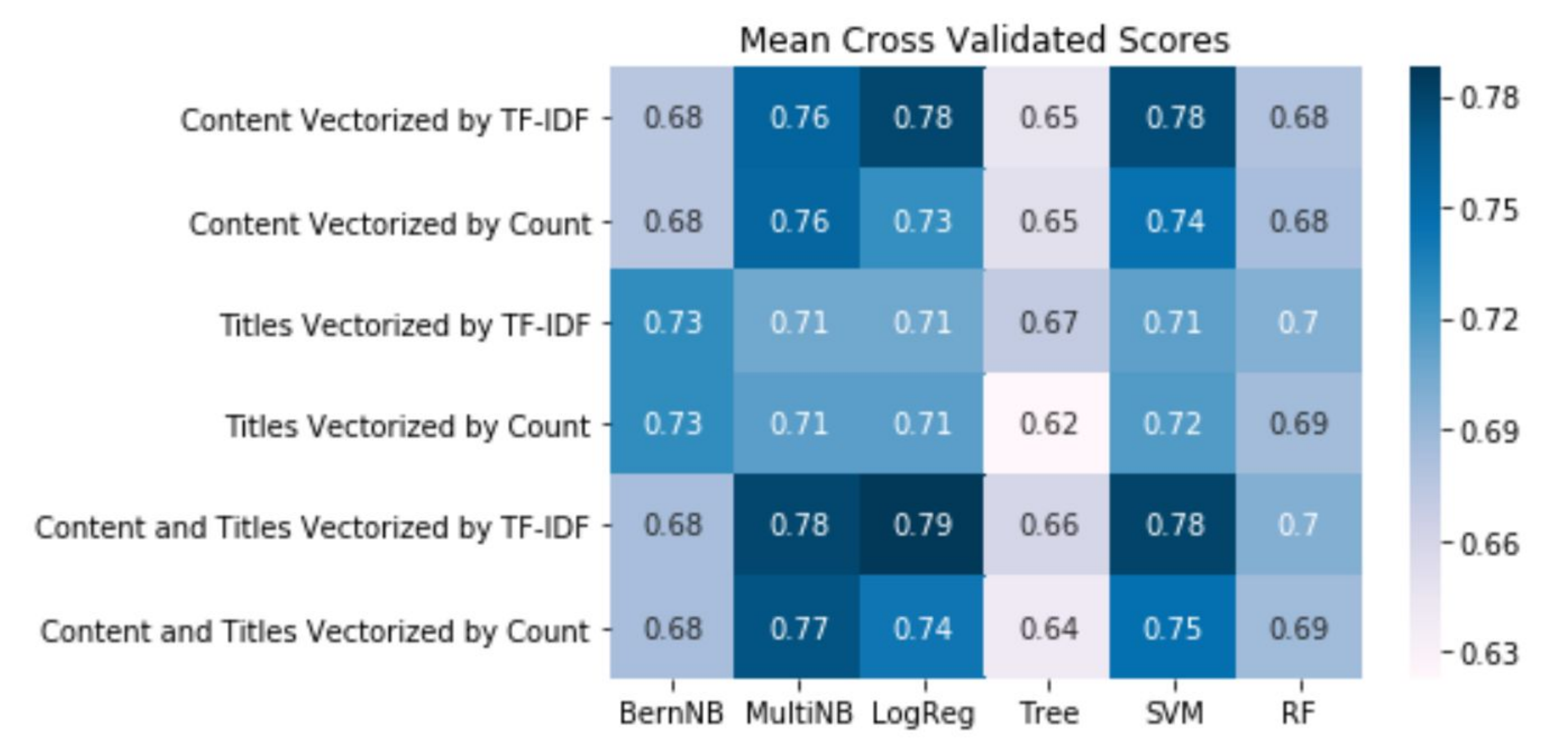


Exploratory Data Analysis

- Length of posts in r/relationship_advice is longer than posts in r/dating_advice on average
- Posts in r/relationship_advice can garner significant larger amounts of interest (e.g. number of comments)
- r/relationship_advice has a more active community (2.3M members) vs r/dating_advice (1.2M members)



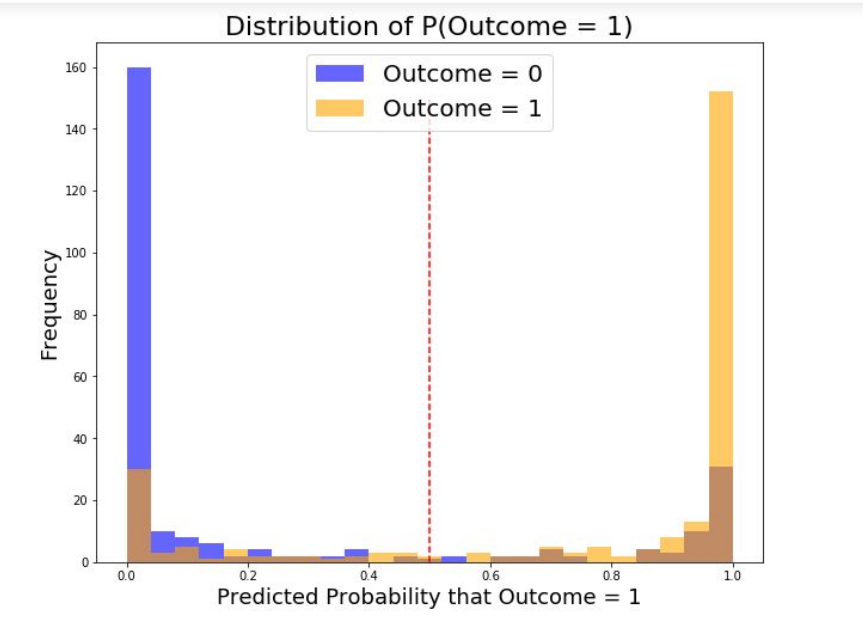
Model Evaluation

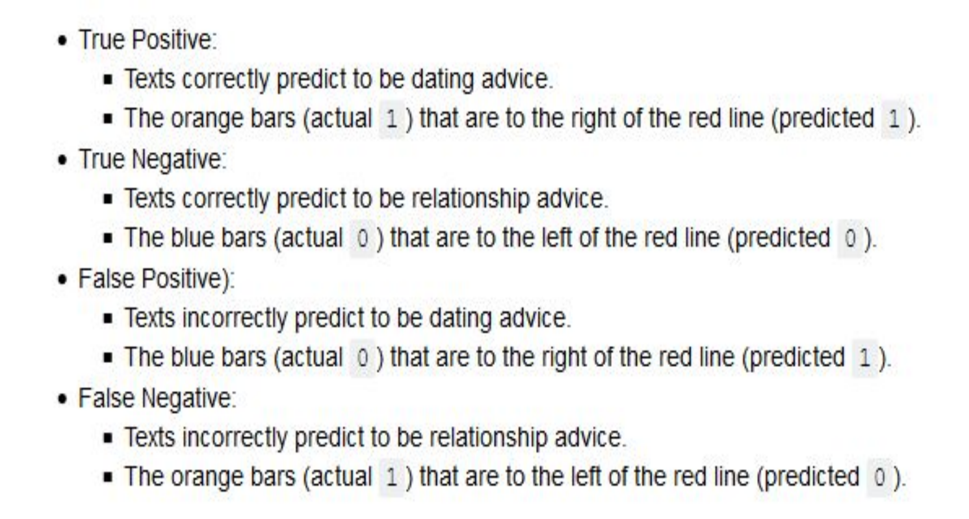


Mean Cross Validated Scores

	BernNB	MultiNB	LogReg	Tree	SVM	RF
Content Vectorized by TF-IDF	0.68	0.76	0.78	0.65	0.78	0.68
Content Vectorized by Count	0.68	0.76	0.73	0.65	0.74	0.68
Titles Vectorized by TF-IDF	0.73	0.71	0.71	0.67	0.71	0.7
Titles Vectorized by Count	0.73	0.71	0.71	0.62	0.72	0.69
Content and Titles Vectorized by TF-IDF	0.68	0.78	0.79	0.66	0.78	0.7
Content and Titles Vectorized by Count	0.68	0.77	0.74	0.64	0.75	0.69

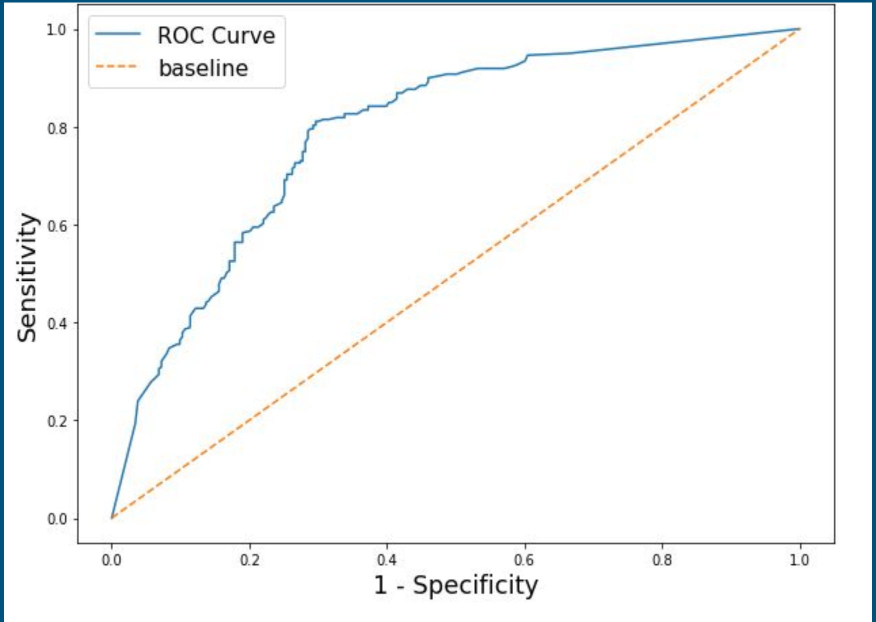
Model Evaluation

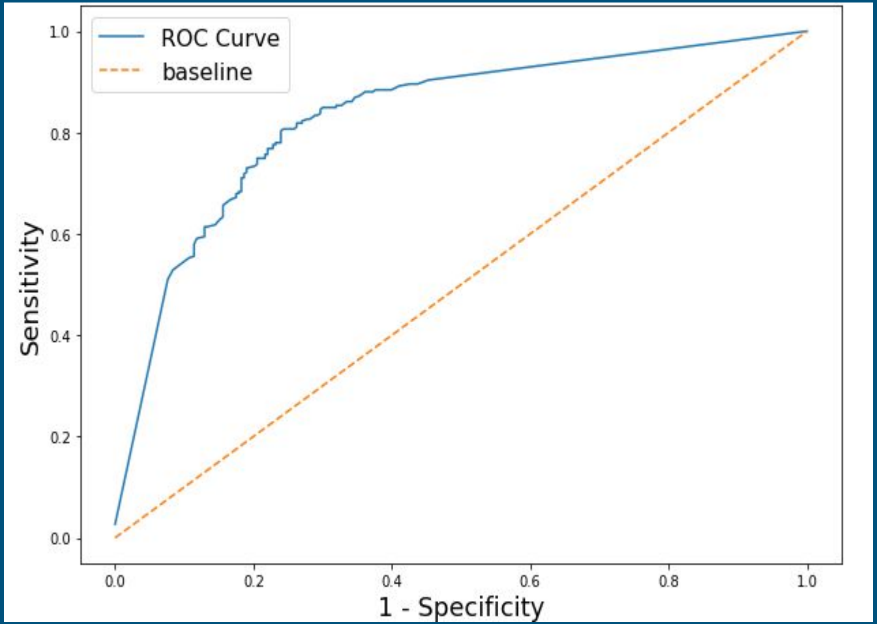




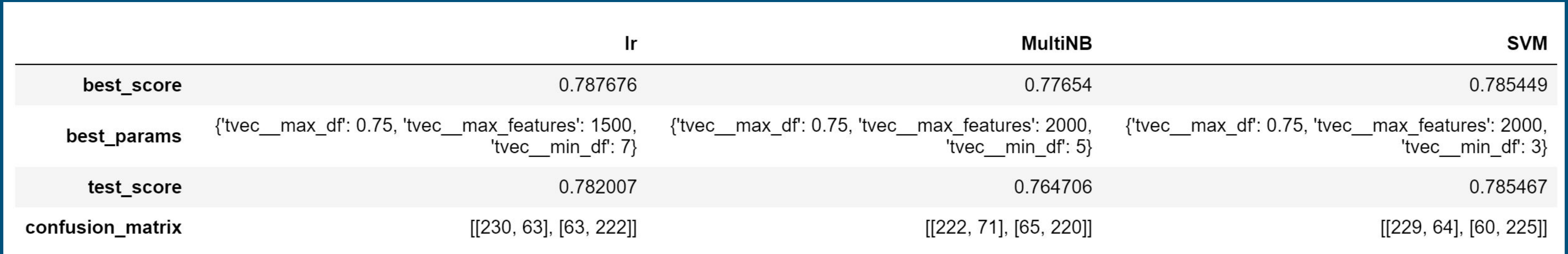
- True Positive:
  - Texts correctly predict to be dating advice.
  - The orange bars (actual `1`) that are to the right of the red line (predicted `1`).
- True Negative:
  - Texts correctly predict to be relationship advice.
  - The blue bars (actual `0`) that are to the left of the red line (predicted `0`).
- False Positive):
  - Texts incorrectly predict to be dating advice.
  - The blue bars (actual `0`) that are to the right of the red line (predicted `1`).
- False Negative:
  - Texts incorrectly predict to be relationship advice.
  - The orange bars (actual `1`) that are to the left of the red line (predicted `0`).

Model Evaluation





Model Evaluation



	lr	MultiNB	SVM
best_score	0.787676	0.77654	0.785449
best_params	{'tvec__max_df': 0.75, 'tvec__max_features': 1500, 'tvec__min_df': 7}	{'tvec__max_df': 0.75, 'tvec__max_features': 2000, 'tvec__min_df': 5}	{'tvec__max_df': 0.75, 'tvec__max_features': 2000, 'tvec__min_df': 3}
test_score	0.782007	0.764706	0.785467
confusion_matrix	[[230, 63], [63, 222]]	[[222, 71], [65, 220]]	[[229, 64], [60, 225]]

- Selected Model: Logistic Regression
- Scores for Logistic Regression and SVM both high, but SVM is a complex model, thus we chose the Logistic Regression model
- When separately fitted outside of the gridsearch, scores suggested SVM had some overfitting

Limitations & Further Improvements/Explorations

- Reddit community is global (but mostly skewed towards the States and a few European countries), so audience/contributors come from diverse backgrounds .
  - Model might not be able to perform as well as it should since it may not be able to pick up local keywords/acronyms such as, for example, BTO, CPF, ROM, etc.
- Extend processing to other models or leverage on other hyperparameters such as ngram = 2 or 3

Limitations & Further Improvements/Explorations

- Seasonality and timing matters for dating/relationship issues!
    - Webscraping at different timings yield slightly different results
    - Seasonality: webscraping during Feb (Valentine's Day) or festive periods (year-end)
- Unable to process videos/images
    - With further refining, can extend model/algorithm to chatbots for dating/relationship consultancies
    - E.g. screenshot of WhatsApp text messages

Conclusion

- Model chosen: TVEC transformation + LogReg
  - Model score: 0.782
  - Baseline score: 0.492 (dating); 0.508 (relationship)
- Many areas for improvement, but model is robust and can be further developed and used for real-life application (semi-automating dating/relationship consultation).

Sources

https://www.straitstimes.com/singapore/3-in-10-here-have-faced-domestic-abuse-cases-poll

https://www.quora.com/What-are-the-advantages-and-disadvantages-of-TF-IDF

https://www.theguardian.com/lifeandstyle/2019/oct/14/cuffing-season-are-people-really-coupling-up-just-because-it-is-winter