# Relationship Advice Classification

Relationship Advice vs Breakups

# CONTENT

1. Overview
2. Problem Statement
3. Data Gathering
4. EDA
5. Cleaning of Data
6. Baseline model
7. Alternative models
8. Key findings
9. Issues with model
10. Conclusions

# Overview

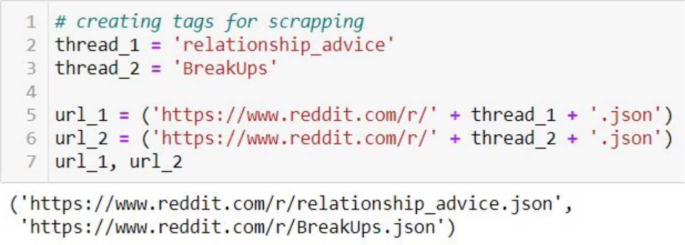
- Relationship advisor
- Classify text as either relationship or breakups
- Data collection of subreddit post  (Relationship_advice & BreakUps)

# Problem Statement

Create a classification model to determine the nature of the text being relationship advice or breakup.

# Data Collection

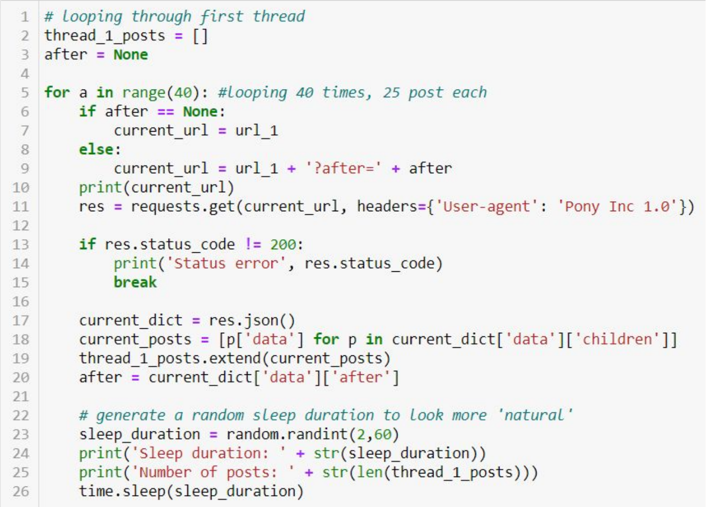
## Setting Thread to webscrap



```python
# creating tags for scrapping
thread_1 = 'relationship_advice'
thread_2 = 'BreakUps'

url_1 = ('https://www.reddit.com/r/' + thread_1 + '.json')
url_2 = ('https://www.reddit.com/r/' + thread_2 + '.json')
url_1, url_2
```

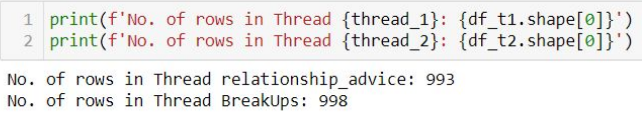
```
('https://www.reddit.com/r/relationship_advice.json',
 'https://www.reddit.com/r/BreakUps.json')
```

## Code for Webscrapping



```python
# looping through first thread
thread_1_posts = []
after = None

for a in range(40): #looping 40 times, 25 post each
    if after == None:
        current_url = url_1
    else:
        current_url = url_1 + '?after=' + after
    print(current_url)
    res = requests.get(current_url, headers={'User-agent': 'Pony Inc 1.0'})

    if res.status_code != 200:
        print('Status error', res.status_code)
        break

    current_dict = res.json()
    current_posts = [p['data'] for p in current_dict['data']['children']]
    thread_1_posts.extend(current_posts)
    after = current_dict['data']['after']

    # generate a random sleep duration to look more 'natural'
    sleep_duration = random.randint(2,60)
    print('Sleep duration: ' + str(sleep_duration))
    print('Number of posts: ' + str(len(thread_1_posts)))
    time.sleep(sleep_duration)
```

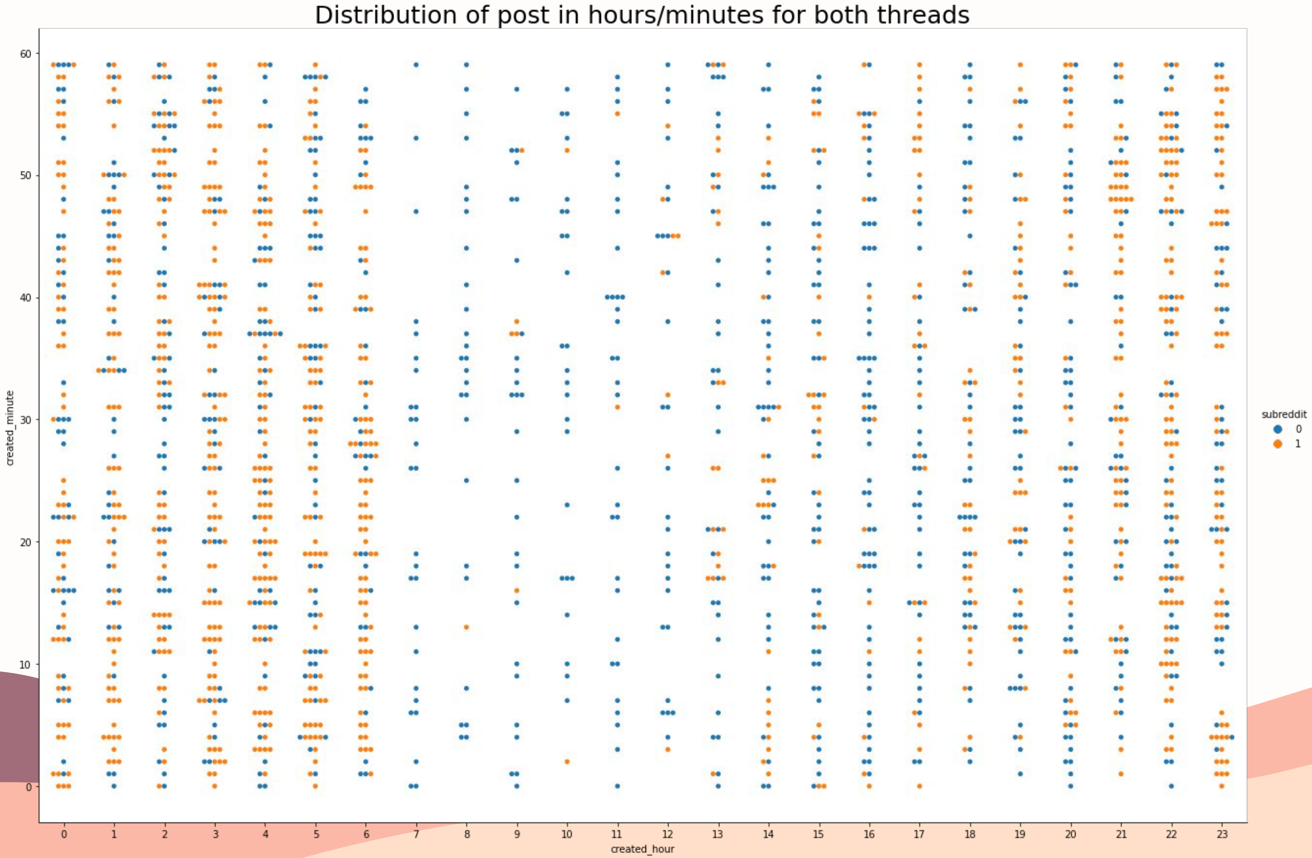
## Amount of text scrapped



```python
print(f'No. of rows in Thread {thread_1}: {df_t1.shape[0]}')
print(f'No. of rows in Thread {thread_2}: {df_t2.shape[0]}')
```

```
No. of rows in Thread relationship_advice: 993
No. of rows in Thread BreakUps: 998
```

# EDA

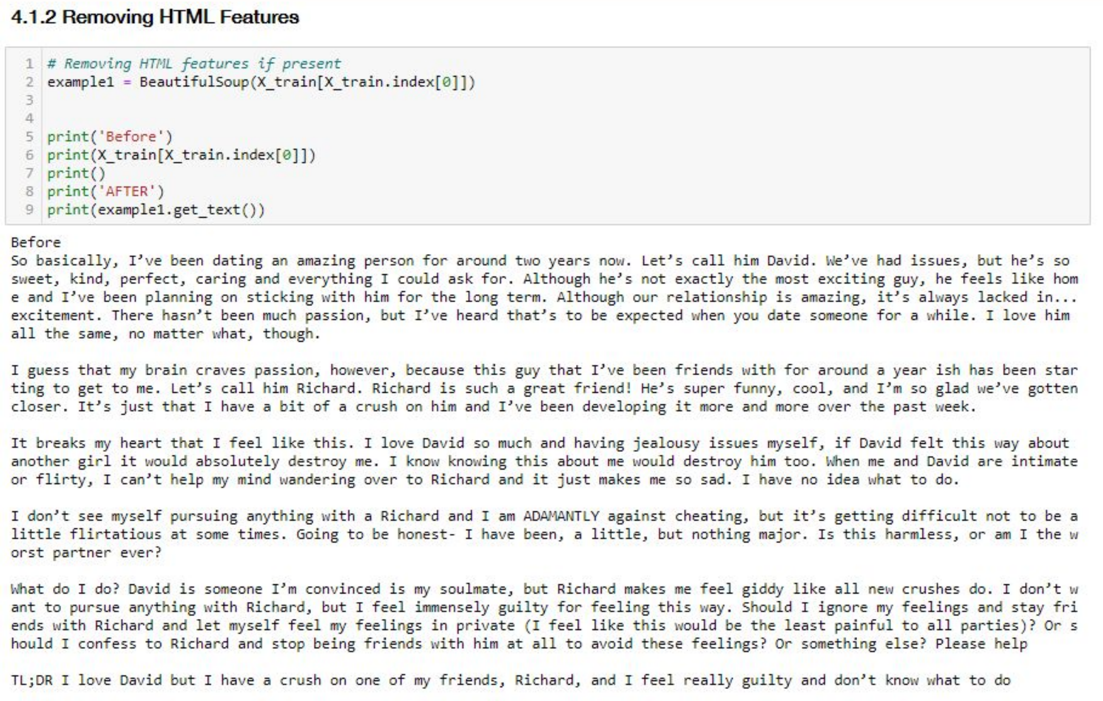


Blue = Breakup

Orange = Relationship advice

# Cleaning of Data

1. **Removing HTML Features**
2. Removing numerical values
3. Lowercase text
4. Removing stopwords
5. Stemming Text



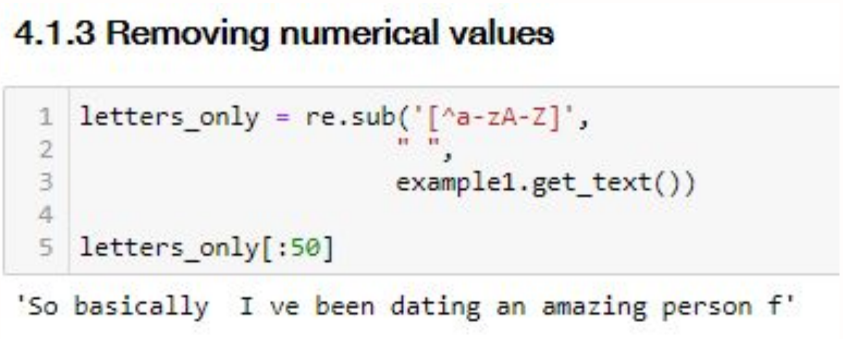
### 4.1.2 Removing HTML Features

```python
# Removing HTML features if present
example1 = BeautifulSoup(X_train[X_train.index[0]])


print('Before')
print(X_train[X_train.index[0]])
print()
print('AFTER')
print(example1.get_text())
```

```
Before
So basically, I’ve been dating an amazing person for around two years now. Let’s call him David. We’ve had issues, but he’s so
sweet, kind, perfect, caring and everything I could ask for. Although he’s not exactly the most exciting guy, he feels like hom
e and I’ve been planning on sticking with him for the long term. Although our relationship is amazing, it’s always lacked in...
excitement. There hasn’t been much passion, but I’ve heard that’s to be expected when you date someone for a while. I love him
all the same, no matter what, though.

I guess that my brain craves passion, however, because this guy that I’ve been friends with for around a year ish has been star
ting to get to me. Let’s call him Richard. Richard is such a great friend! He’s super funny, cool, and I’m so glad we’ve gotten
closer. It’s just that I have a bit of a crush on him and I’ve been developing it more and more over the past week.

It breaks my heart that I feel like this. I love David so much and having jealousy issues myself, if David felt this way about
another girl it would absolutely destroy me. I know knowing this about me would destroy him too. When me and David are intimate
or flirty, I can’t help my mind wandering over to Richard and it just makes me so sad. I have no idea what to do.

I don’t see myself pursuing anything with a Richard and I am ADAMANTLY against cheating, but it’s getting difficult not to be a
little flirtatious at some times. Going to be honest- I have been, a little, but nothing major. Is this harmless, or am I the w
orst partner ever?

What do I do? David is someone I’m convinced is my soulmate, but Richard makes me feel giddy like all new crushes do. I don’t w
ant to pursue anything with Richard, but I feel immensely guilty for feeling this way. Should I ignore my feelings and stay fri
ends with Richard and let myself feel my feelings in private (I feel like this would be the least painful to all parties)? Or s
hould I confess to Richard and stop being friends with him at all to avoid these feelings? Or something else? Please help

TL;DR I love David but I have a crush on one of my friends, Richard, and I feel really guilty and don’t know what to do
```

# Cleaning of Data

1. Removing HTML Features
2. Removing numerical values
3. Lowercase text
4. Removing stopwords
5. Stemming Text



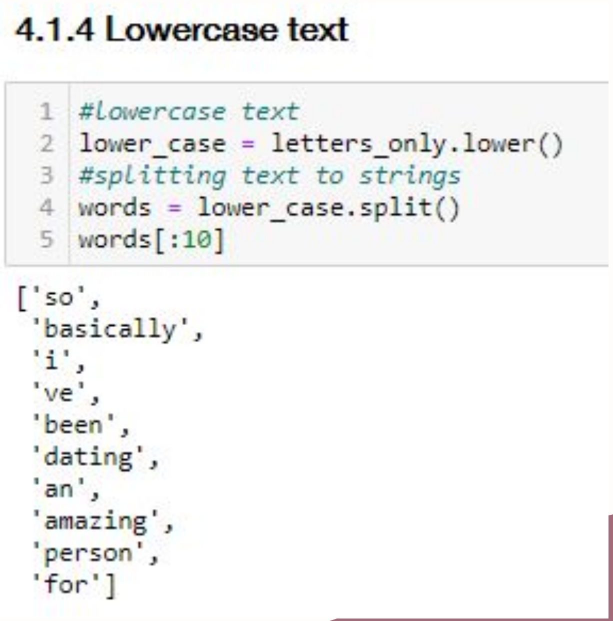
### 4.1.3 Removing numerical values

```
letters_only = re.sub('[^a-zA-Z]',
                      " ",
                      example1.get_text())

letters_only[:50]
```

```
'So basically  I ve been dating an amazing person f'
```

# Cleaning of Data

1. Removing HTML Features
2. Removing numerical values
3. Lowercase text
4. Removing stopwords
5. Stemming Text



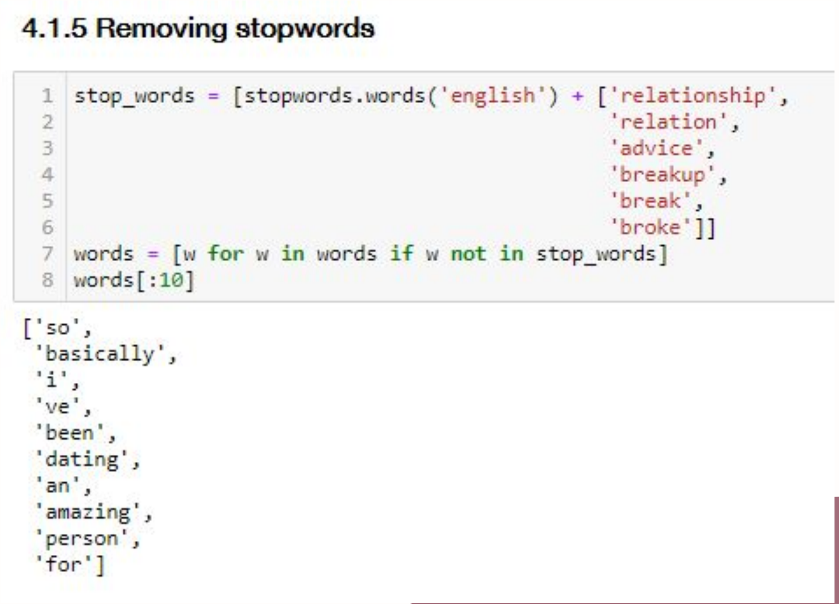
### 4.1.4 Lowercase text

```
#lowercase text
lower_case = letters_only.lower()
#splitting text to strings
words = lower_case.split()
words[:10]
```

```
['so',
 'basically',
 'i',
 've',
 'been',
 'dating',
 'an',
 'amazing',
 'person',
 'for']
```

# Cleaning of Data

1. Removing HTML Features
2. Removing numerical values
3. Lowercase text
4. **Removing stopwords**
5. Stemming Text



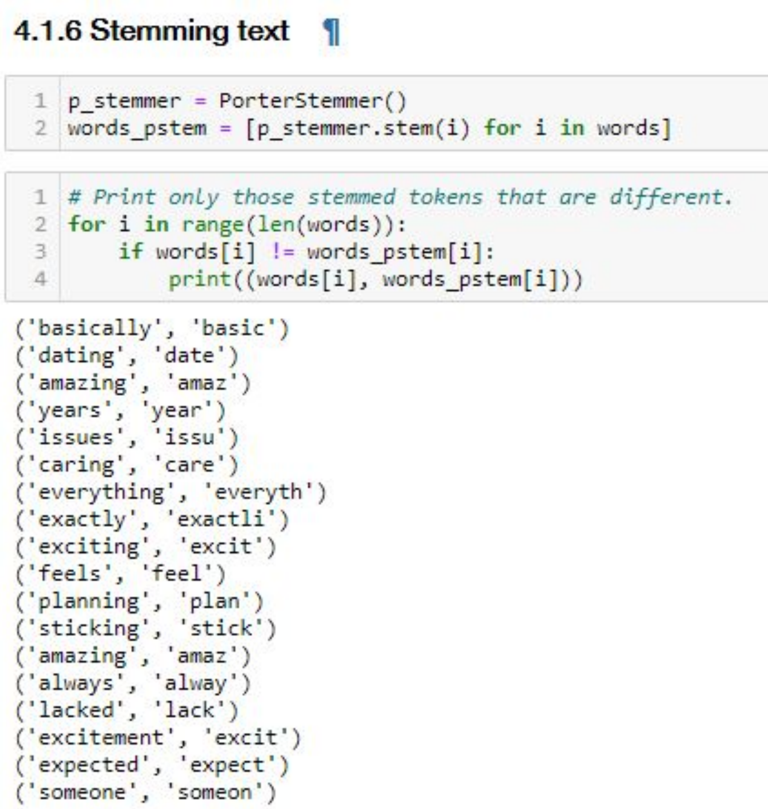
### 4.1.5 Removing stopwords

```
stop_words = [stopwords.words('english') + ['relationship',
                                              'relation',
                                              'advice',
                                              'breakup',
                                              'break',
                                              'broke']]
words = [w for w in words if w not in stop_words]
words[:10]
```

```
['so',
 'basically',
 'i',
 've',
 'been',
 'dating',
 'an',
 'amazing',
 'person',
 'for']
```

# Cleaning of Data

1. Removing HTML Features
2. Removing numerical values
3. Lowercase text
4. Removing stopwords
5. **Stemming Text**



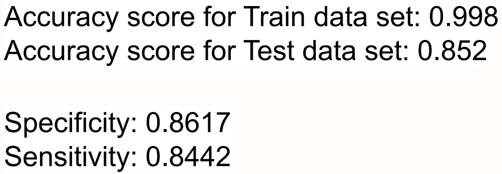
### 4.1.6 Stemming text ¶

```
p_stemmer = PorterStemmer()
words_pstem = [p_stemmer.stem(i) for i in words]
```

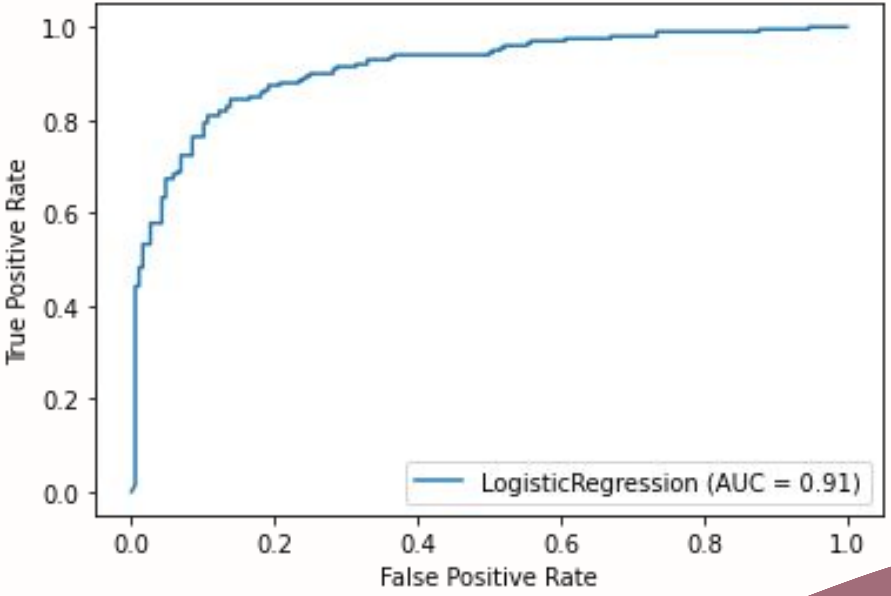
```
# Print only those stemmed tokens that are different.
for i in range(len(words)):
    if words[i] != words_pstem[i]:
        print((words[i], words_pstem[i]))
```

```
('basically', 'basic')
('dating', 'date')
('amazing', 'amaz')
('years', 'year')
('issues', 'issu')
('caring', 'care')
('everything', 'everyth')
('exactly', 'exactli')
('exciting', 'excit')
('feels', 'feel')
('planning', 'plan')
('sticking', 'stick')
('amazing', 'amaz')
('always', 'alway')
('lacked', 'lack')
('excitement', 'excit')
('expected', 'expect')
('someone', 'someon')
```

# Baseline Model (Linear Regression)

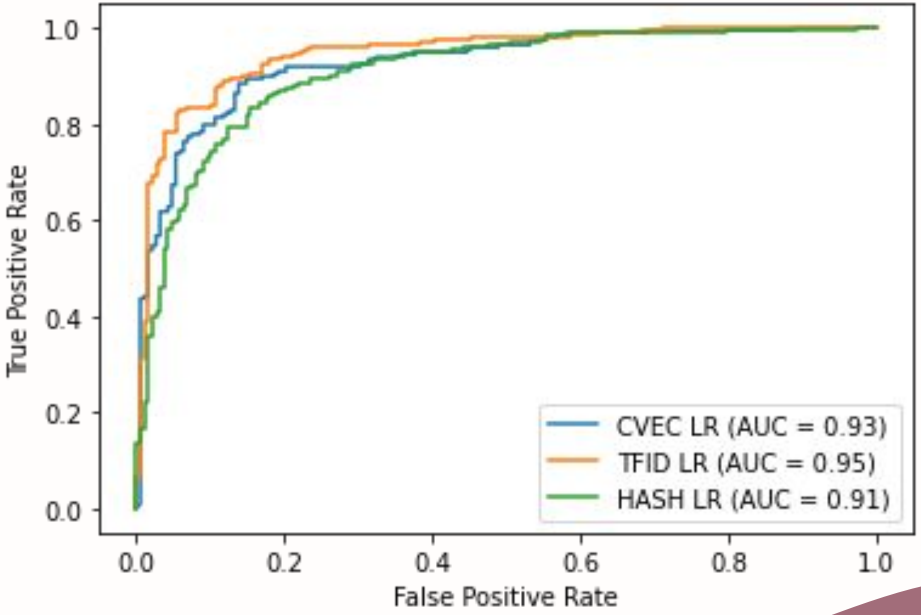


```
Accuracy score for Train data set: 0.998
Accuracy score for Test data set: 0.852

Specificity: 0.8617
Sensitivity: 0.8442
```



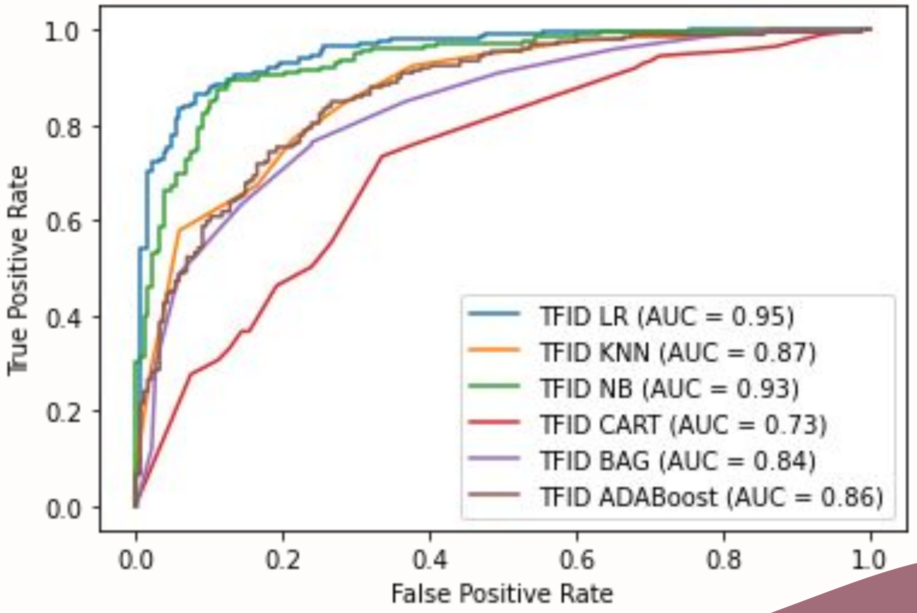
# Alternative Model (part 1)

```
CVEC LogReg Train Score: 0.8134
CVEC LogReg Test Score: 0.8656
---
TFID LogReg Train Score: 0.8542
TFID LogReg Test Score: 0.8759
---
HASH LogReg Train Score: 0.8050
HASH LogReg Test Score: 0.8346
```



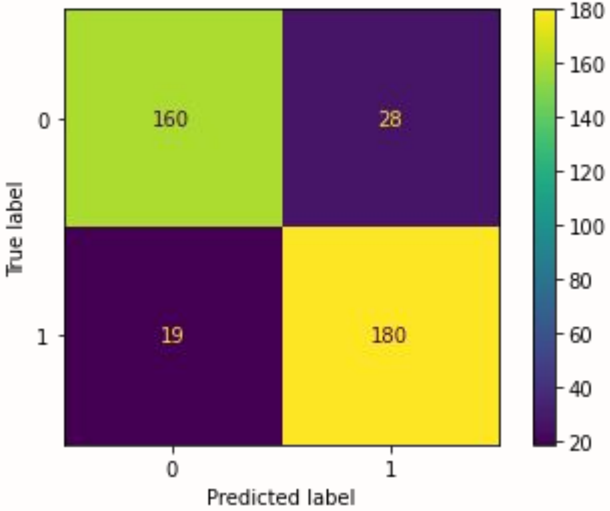
# Alternative Model (part 2)

```
TFID LogReg Train Score: 0.8581
TFID LogReg Test Score: 0.8785
---
TFID KNN Train Score: 0.7953
TFID KNN Test Score: 0.7829
---
TFID NB Train Score: 0.8354
TFID NB Test Score: 0.8501
---
TFID CART Train Score: 0.6819
TFID CART Test Score: 0.7002
---
TFID BAG Train Score: 0.7661
TFID BAG Test Score: 0.7622
---
TFID ADABoost Train Score: 0.7778
TFID ADABoost Test Score: 0.7777
```

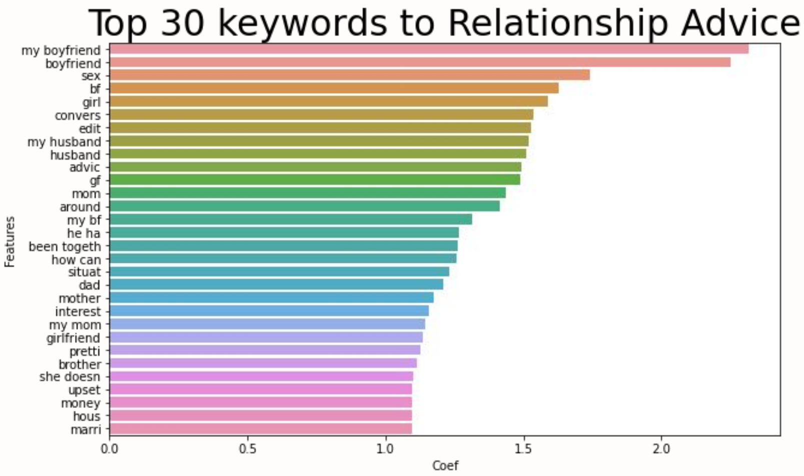


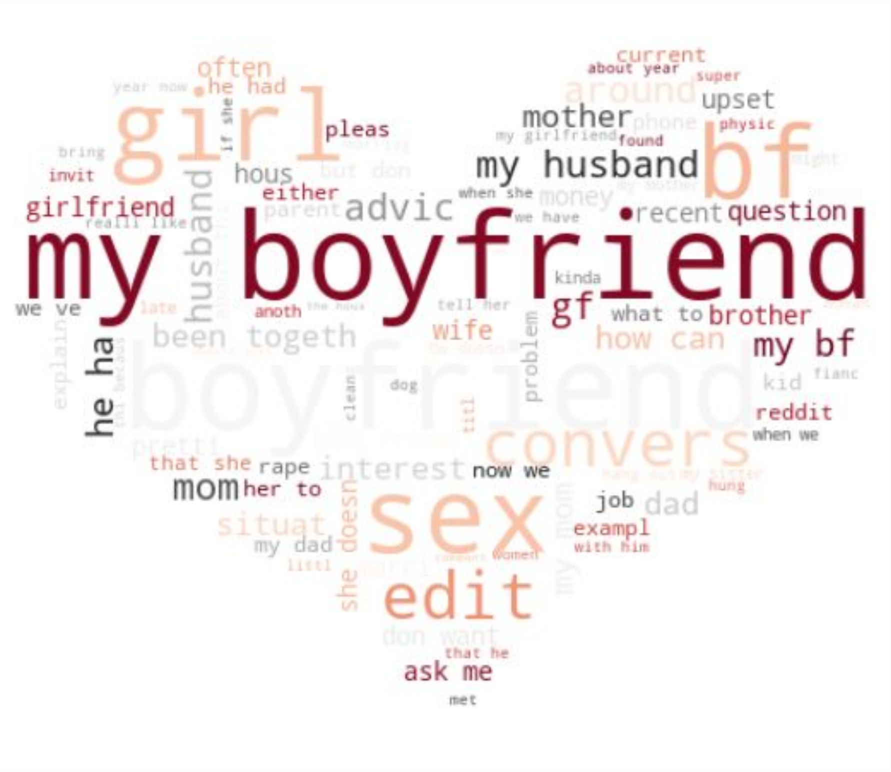
# Final Model (TFID Linear Regression)

True Negatives: 160
False Positives: 28
False Negatives: 19
True Positives: 180
Sensitivity, Accuracy of Thread 1: 0.9045
Specificity, Accuracy of Thread 2: 0.8511

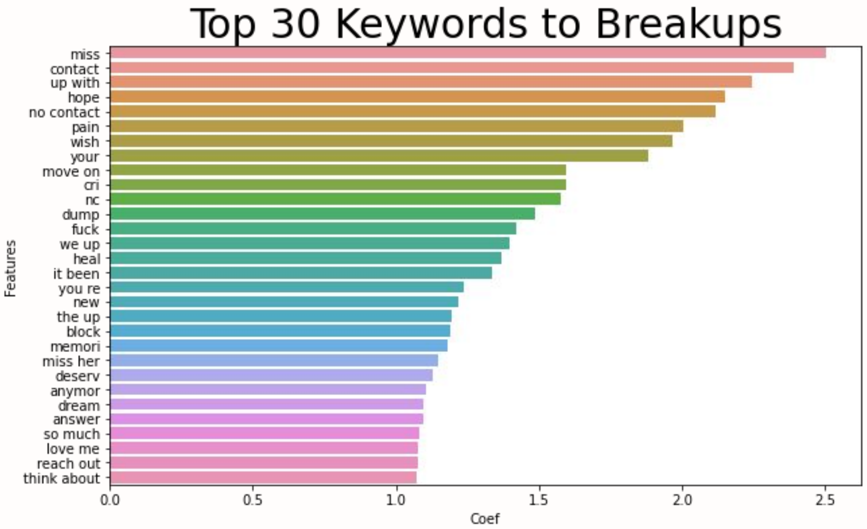


# Key findings:
# Relationship advice

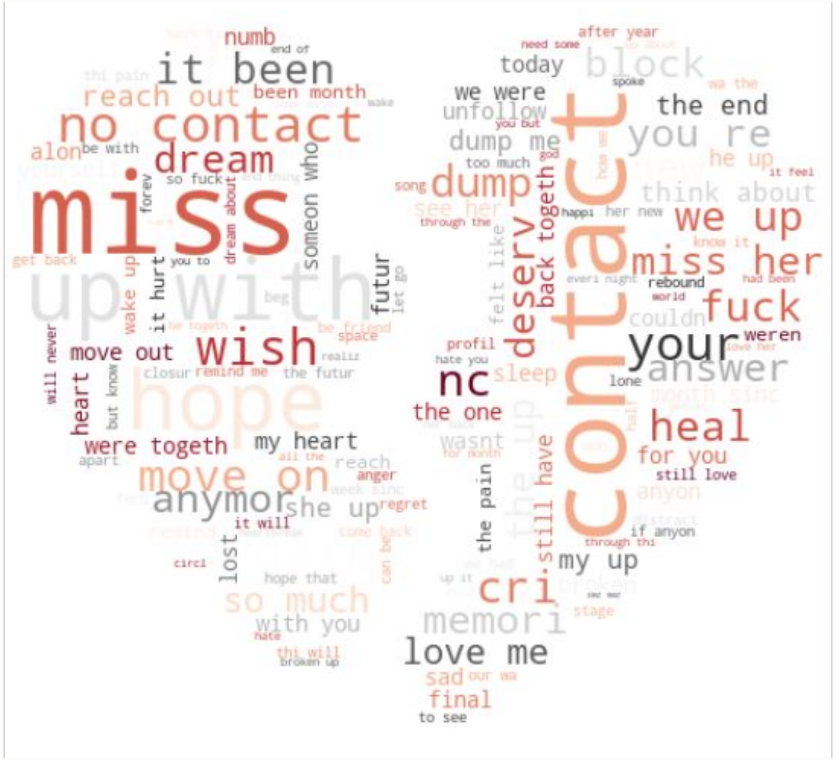

Top 30 keywords to Relationship Advice



# Key findings: Breakup

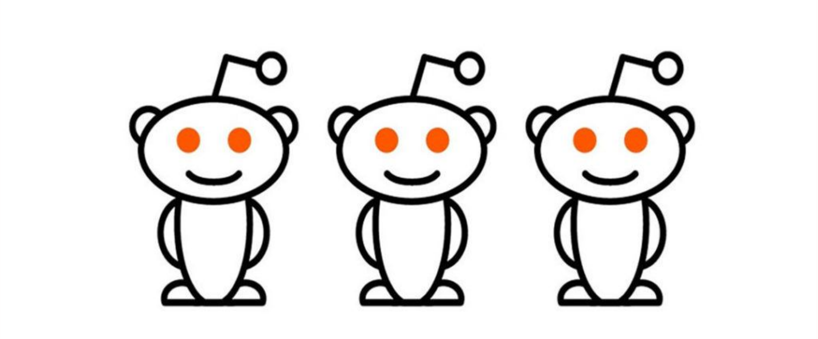


Top 30 Keywords to Breakups



# Limitation of model

Misclassification of
user posted Subreddit



Limitation of hardware
Computational power

# Conlusion

- Webscrapping Subreddit
  - Relationship_advice
  - BreakUps
- Data Cleaning/ Exploration
- Natural Language Exploration
  - Stopword, Stemming, HTML, numerical values
- Modelling
  - Classification models
- Evaluation
  - Linear Regression model