

# Relationship Advice Classification

Relationship Advice vs Breakups




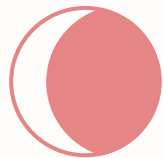
# CONTENT

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

```
1 # creating tags for scrapping
2 thread_1 = 'relationship_advice'
3 thread_2 = 'BreakUps'
4
5 url_1 = ('https://www.reddit.com/r/' + thread_1 + '.json')
6 url_2 = ('https://www.reddit.com/r/' + thread_2 + '.json')
7 url_1, url_2
```

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

## Amount of text scrapped

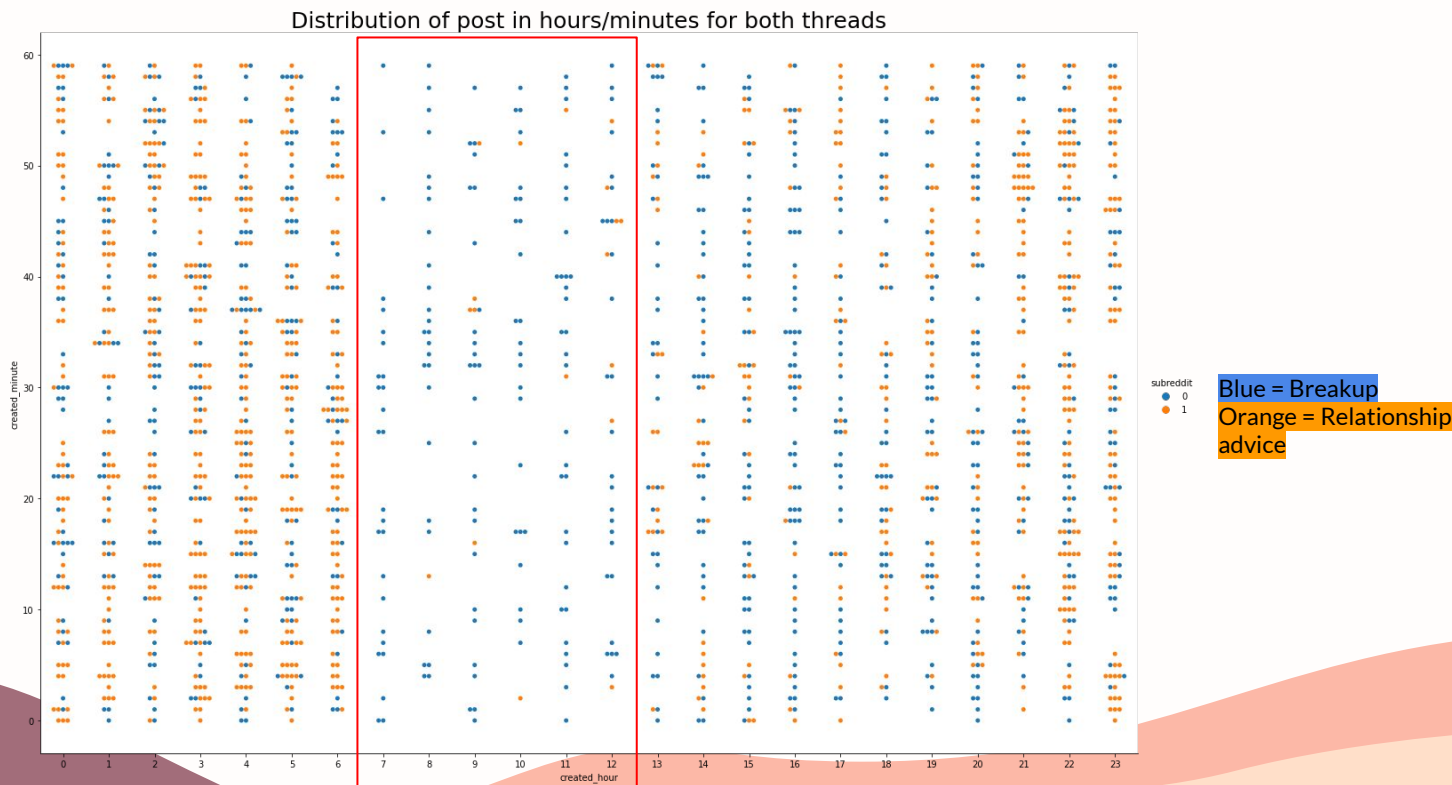
```
1 print(f'No. of rows in Thread {thread_1}: {df_t1.shape[0]}')
2 print(f'No. of rows in Thread {thread_2}: {df_t2.shape[0]}')
```

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

## Code for Webscrapping

```
1 # Looping through first thread
2 thread_1_posts = []
3 after = None
4
5 for a in range(40): #Looping 40 times, 25 post each
6     if after == None:
7         current_url = url_1
8     else:
9         current_url = url_1 + '?after=' + after
10    print(current_url)
11    res = requests.get(current_url, headers={'User-agent': 'Pony Inc 1.0'})
12
13    if res.status_code != 200:
14        print('Status error', res.status_code)
15        break
16
17    current_dict = res.json()
18    current_posts = [p['data'] for p in current_dict['data']['children']]
19    thread_1_posts.extend(current_posts)
20    after = current_dict['data']['after']
21
22    # generate a random sleep duration to look more 'natural'
23    sleep_duration = random.randint(2,60)
24    print('Sleep duration: ' + str(sleep_duration))
25    print('Number of posts: ' + str(len(thread_1_posts)))
26    time.sleep(sleep_duration)
```

# EDA



# 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

```
1 # Removing HTML features if present
2 example1 = BeautifulSoup(X_train[X_train.index[0]])
3
4
5 print('Before')
6 print(X_train[X_train.index[0]])
7 print()
8 print('AFTER')
9 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 home 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 starting 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 worst 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 want to pursue anything with Richard, but I feel immensely guilty for feeling this way. Should I ignore my feelings and stay friends with Richard and let myself feel my feelings in private (I feel like this would be the least painful to all parties)? Or should 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

```
1 letters_only = re.sub('[^a-zA-Z]',  
2                       " ",  
3                       example1.get_text())  
4  
5 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

```
1 #lowercase text
2 lower_case = letters_only.lower()
3 #splitting text to strings
4 words = lower_case.split()
5 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

```
1 stop_words = [stopwords.words('english') + ['relationship',
2                                                    'relation',
3                                                    'advice',
4                                                    'breakup',
5                                                    'break',
6                                                    'broke']]
7 words = [w for w in words if w not in stop_words]
8 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 ¶

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

1 # Print only those stemmed tokens that are different.
2 for i in range(len(words)):
3     if words[i] != words_pstem[i]:
4         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')
```

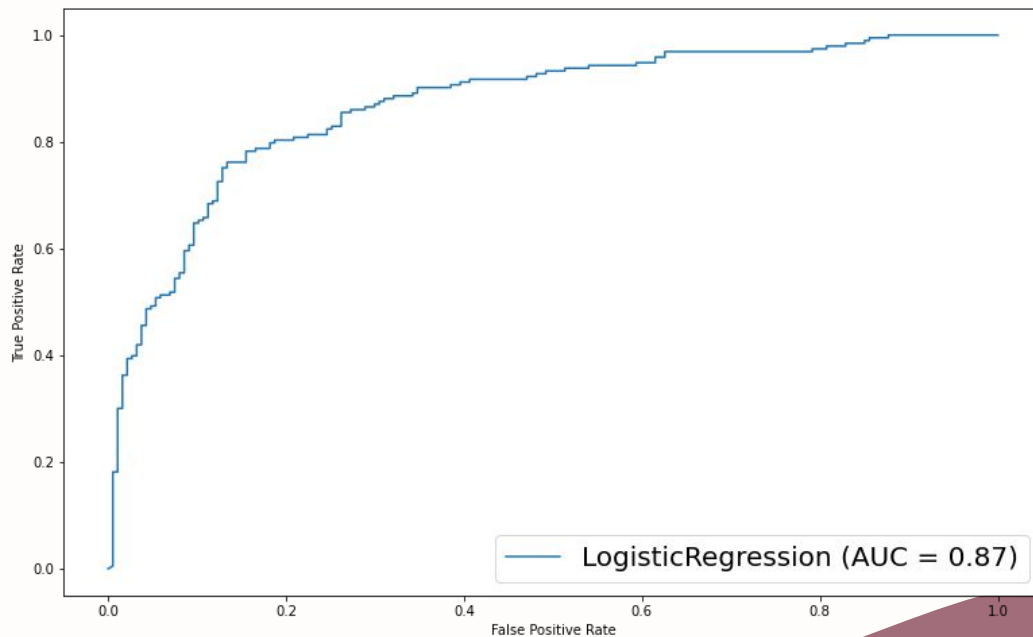
# Baseline Model (Linear Regression)

Accuracy score for Train data set: 0.997

Accuracy score for Test data set: 0.8

Specificity: 0.7968

Sensitivity: 0.8031



# Alternative Model (part 1)

CVEC LogReg Train Score: 0.8161

CVEC LogReg Test Score: 0.7921

---

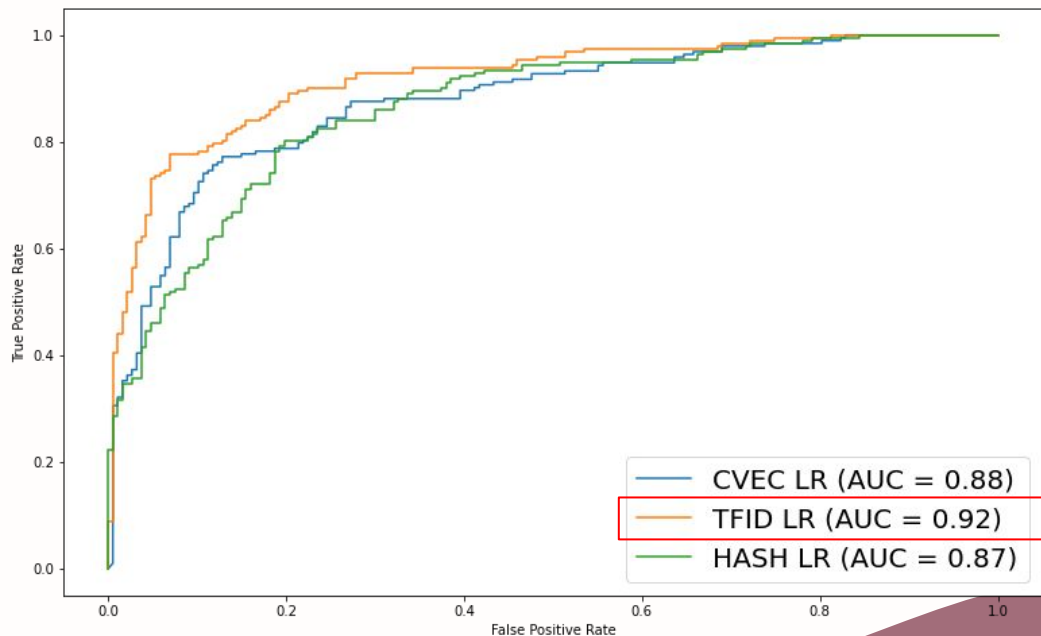
TFID LogReg Train Score: **0.8589**

TFID LogReg Test Score: **0.8394**

---

HASH LogReg Train Score: 0.8069

HASH LogReg Test Score: 0.7894



# Alternative Model (part 2)

TFID LogReg Train Score: **0.8563**

TFID LogReg Test Score: **0.85**

---

TFID KNN Train Score: 0.7924

TFID KNN Test Score: 0.8026

---

TFID NB Train Score: 0.8510

TFID NB Test Score: 0.8289

---

TFID CART Train Score: 0.7055

TFID CART Test Score: 0.6815

---

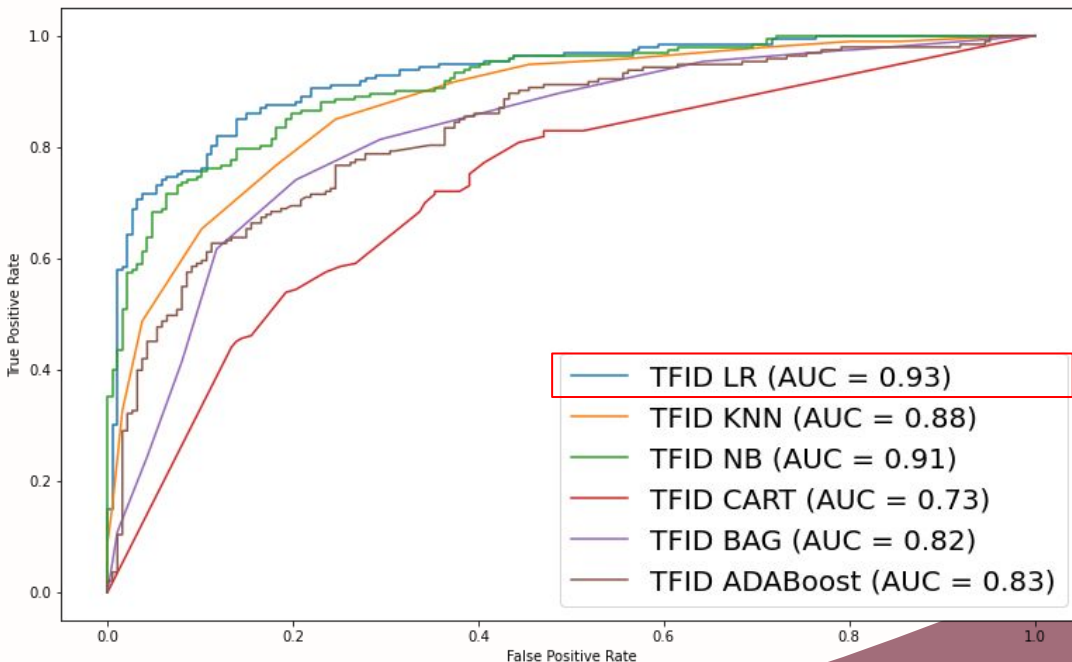
TFID BAG Train Score: 0.7647

TFID BAG Test Score: 0.7684

---

TFID ADABoost Train Score: 0.7944

TFID ADABoost Test Score: 0.7552



# Final Model (TFID Linear Regression)

True Negatives: 156

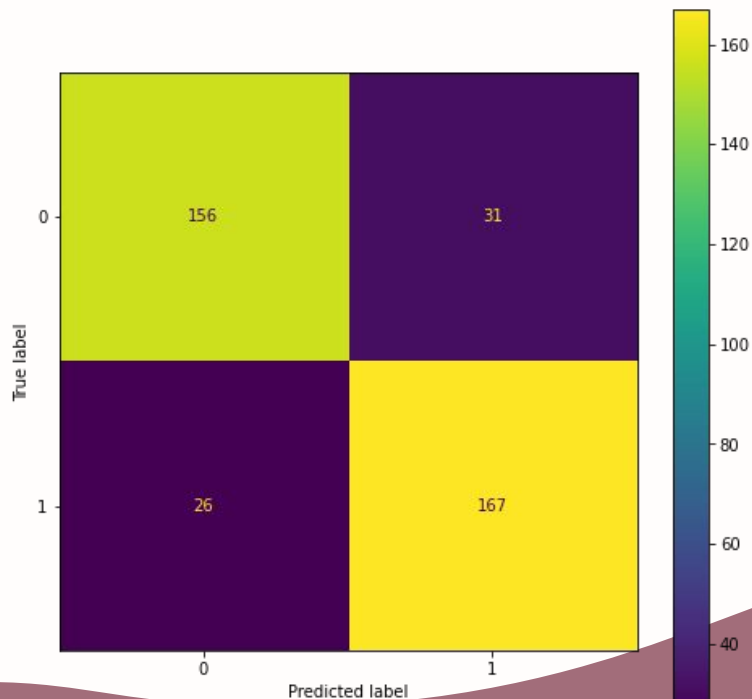
False Positives: 31

False Negatives: 26

True Positives: 167

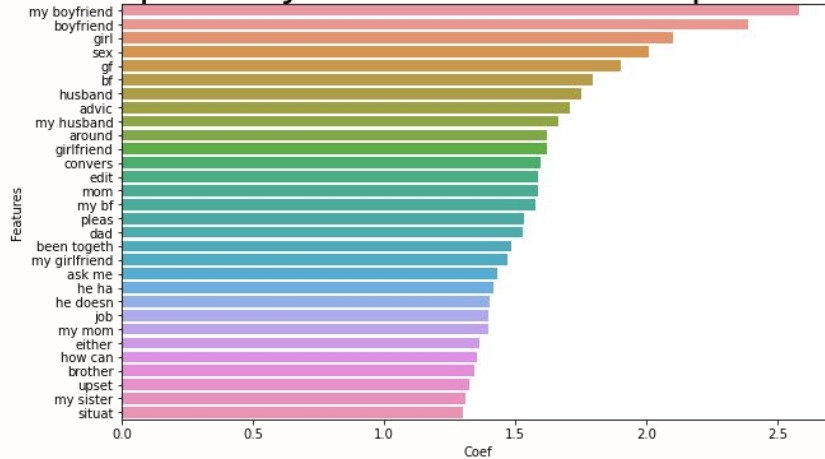
Sensitivity, Accuracy of Thread 1: 0.8653

Specificity, Accuracy of Thread 2: 0.8342



## Key findings: Relationship advice

Top 30 keywords to Relationship Advice

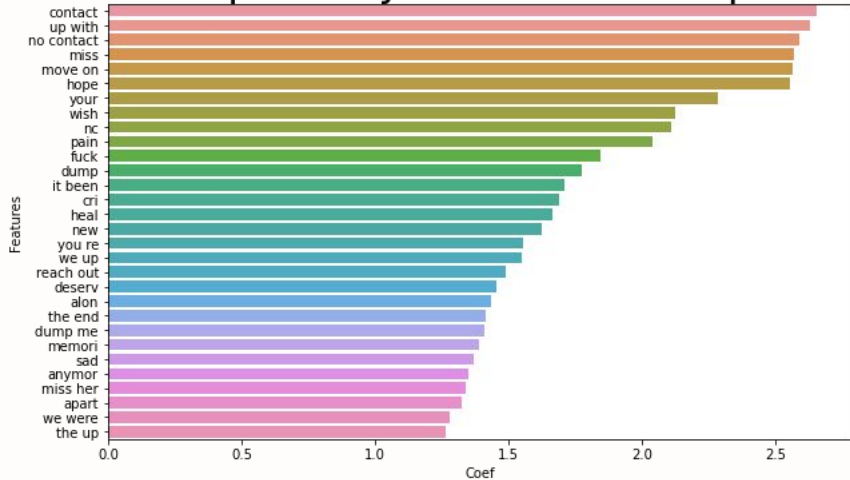




## Key findings:

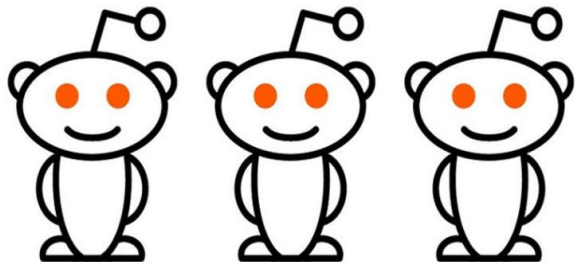
# Breakup

## Top 30 Keywords to Breakups

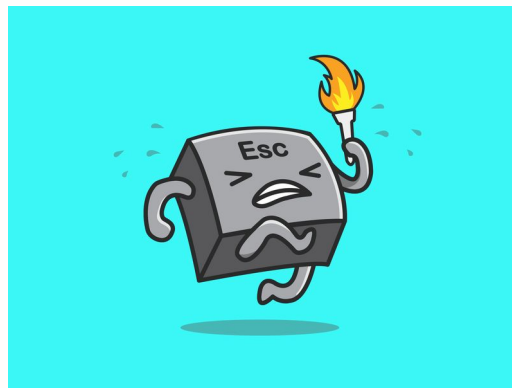


# Limitation of model

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