Relationship Advice Classification

Relationship Advice vs Breakups

CONTENT

Overview

- Make up or Break up
- Creating classification model
- Identify key features of both subreddits

Problem Statement

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

Data Collection

Setting Thread to webscrap

```
# 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

```
# 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)
```

Check and remove duplicates

```
print(f'No.of duplciated post in Thread {thread_1}: {df_t1.selftext.duplicated().sum()}')
print(f'No.of duplciated post in Thread {thread_2}: {df_t2.selftext.duplicated().sum()}')
```

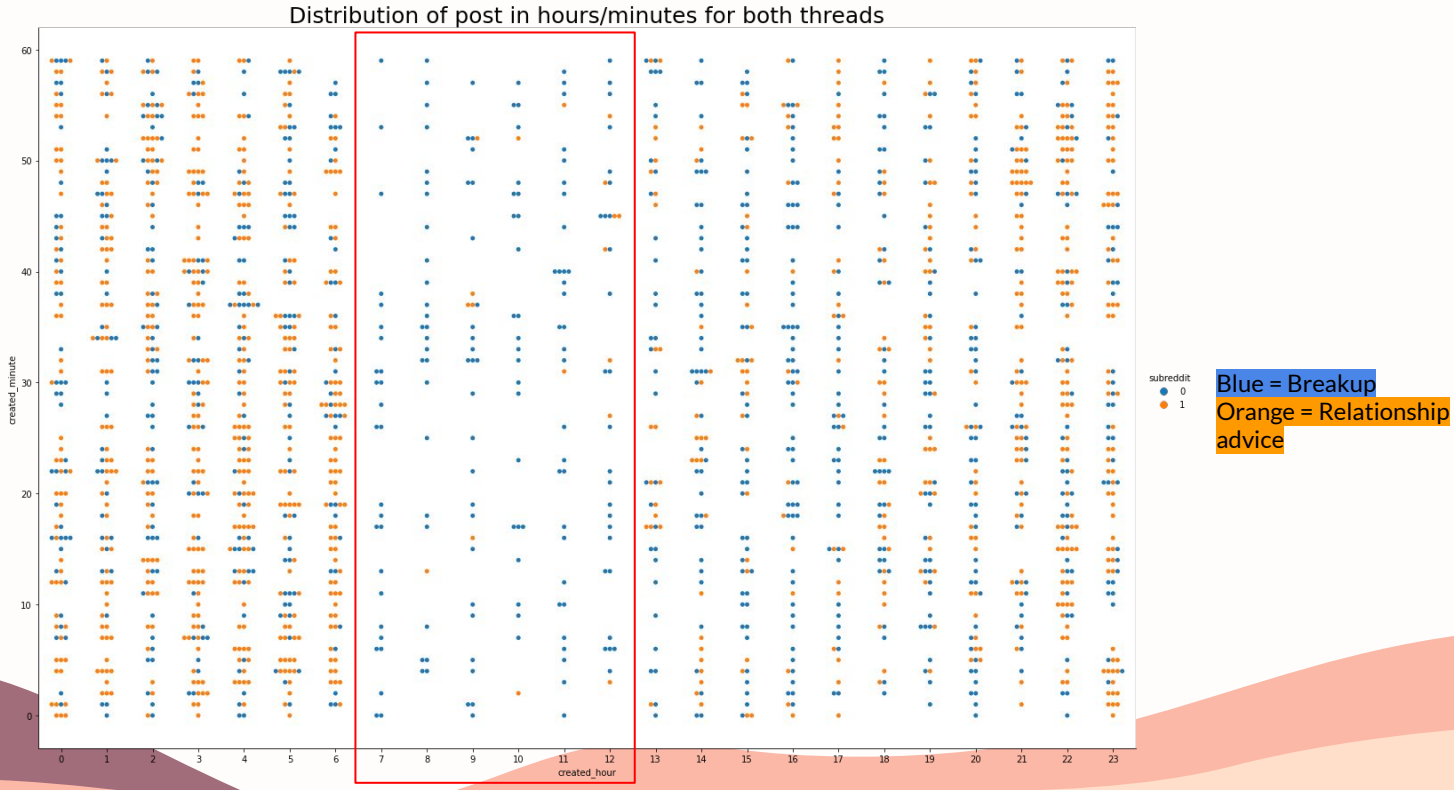
```
No.of duplciated post in Thread relationship_advice: 29
No.of duplciated post in Thread BreakUps: 63
```

Amount of text scrapped

```
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: 964
No. of rows in Thread BreakUps: 935
```

EDA



Blue = Breakup
Orange = Relationship advice

Lesser Relationship advice post

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

```
# 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]
```

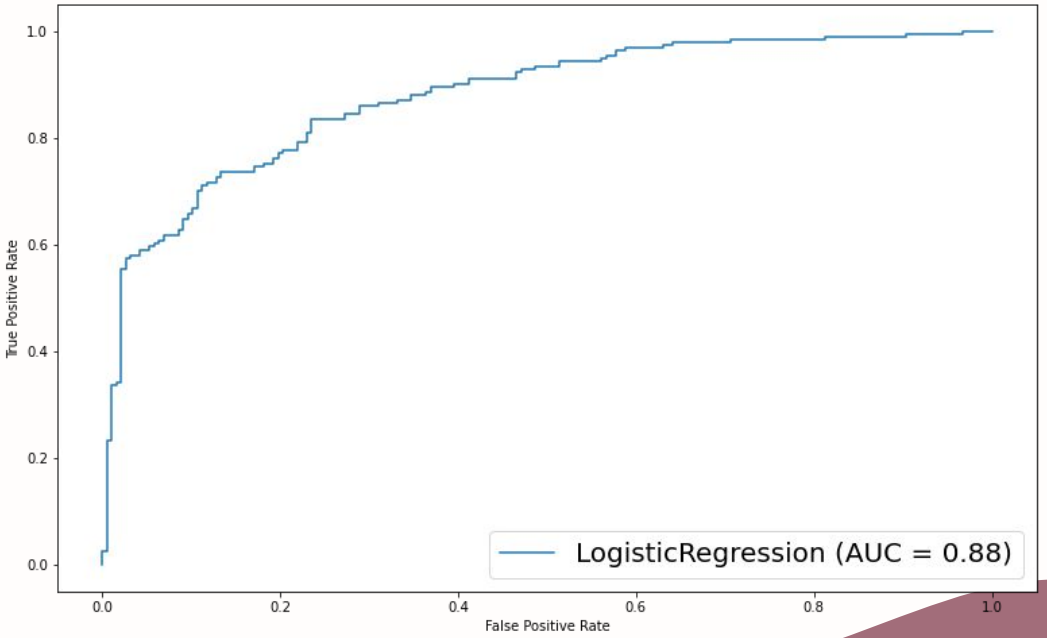
```
# Print only those stemmed tokens that are different.
for i in range(len(words)):
    if words[i] != words_pstem[i]:
        print((words[i], words_pstem[i]))
```

```
('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)

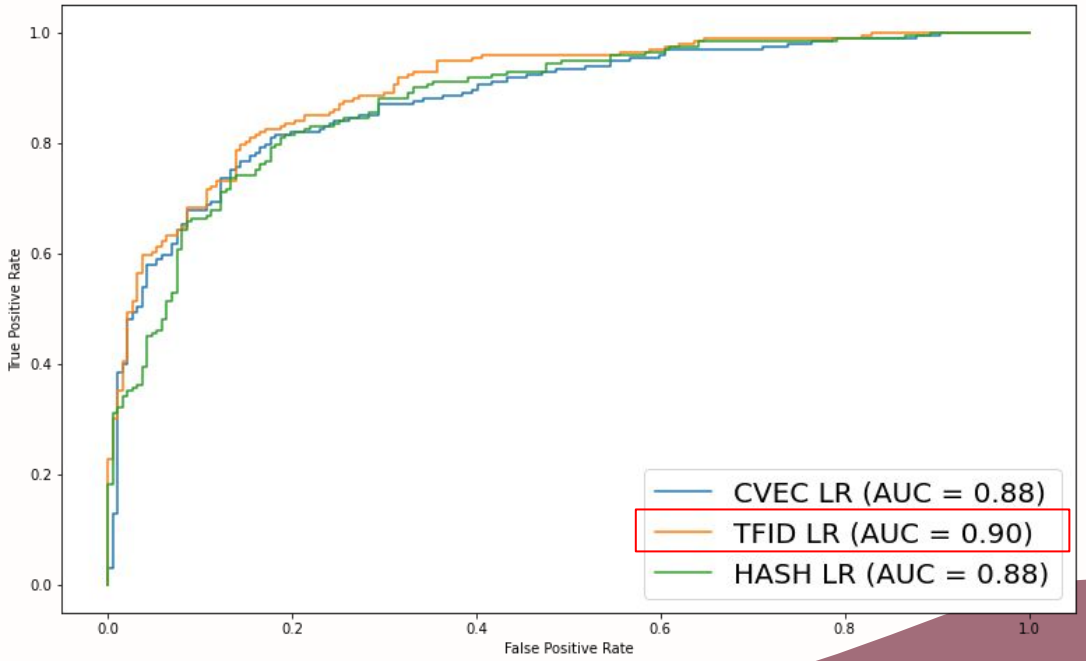
Accuracy score for Train data set: 0.994
Accuracy score for Test data set: 0.781

Specificity: 0.775
Sensitivity: 0.787



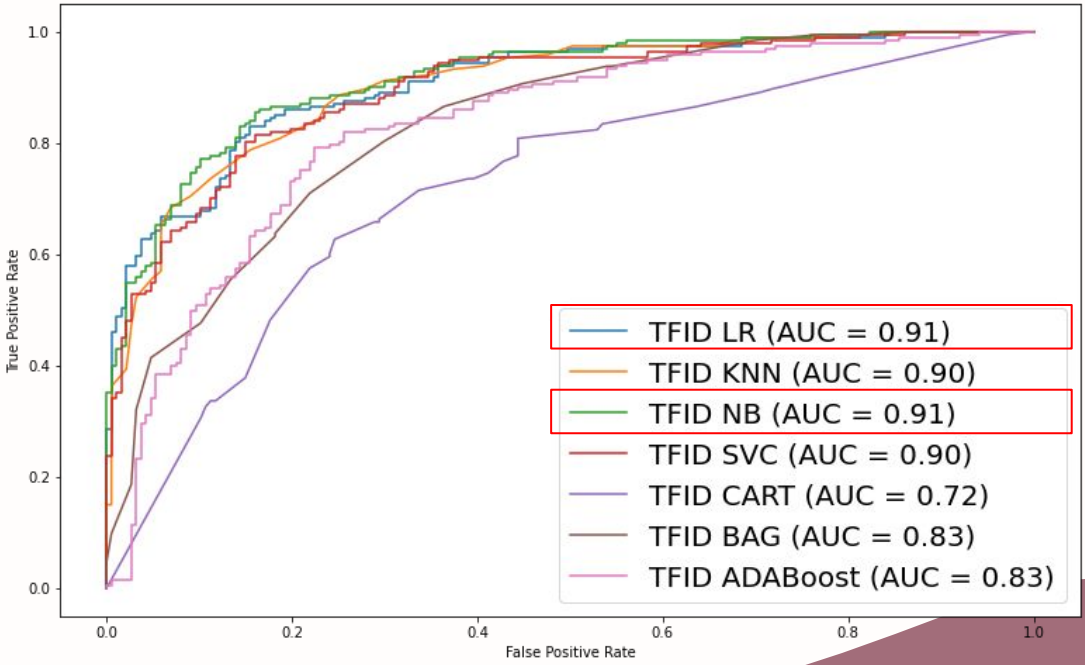
Alternative Model (part 1)

model	vectorizer	train_score	test_score	time(s)
LogisticRegression	CountVectorizer	0.815963	0.810526	35
LogisticRegression	TFIDVectorizer	0.852886	0.807895	25
LogisticRegression	HashingVectorizer	0.821880	0.786842	3

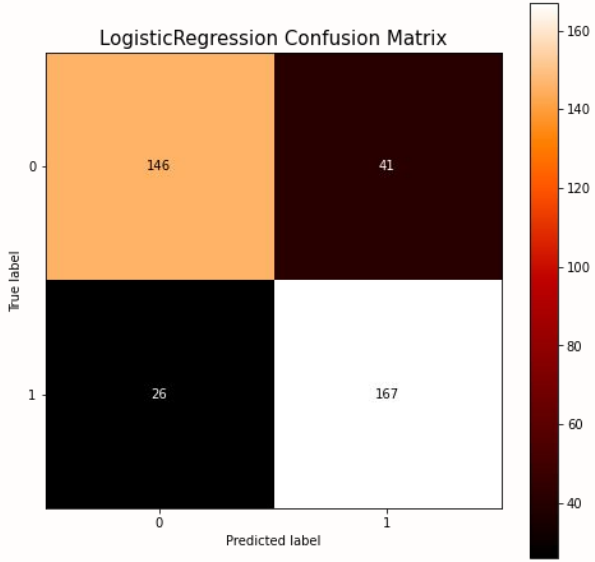


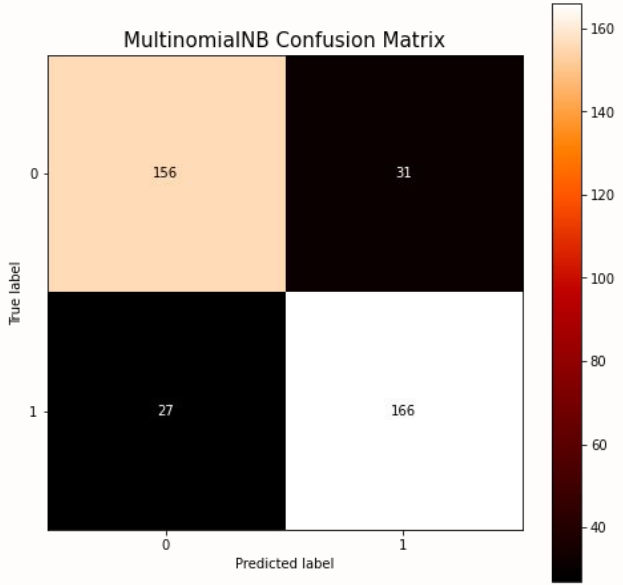
Alternative Model (part 2)

model	vectorizer	train_score	test_score	time(s)
LogisticRegression	TFIDVectorizer	0.848940	0.823684	138
KNN	TFIDVectorizer	0.830465	0.818421	72
MultinomialNB	TFIDVectorizer	0.849603	0.847368	44
SVC	TFIDVectorizer	0.851555	0.805263	73
CART	TFIDVectorizer	0.704468	0.671053	78
BaggingClassifier	TFIDVectorizer	0.773736	0.744737	164
ADABoost	TFIDVectorizer	0.777725	0.784211	384



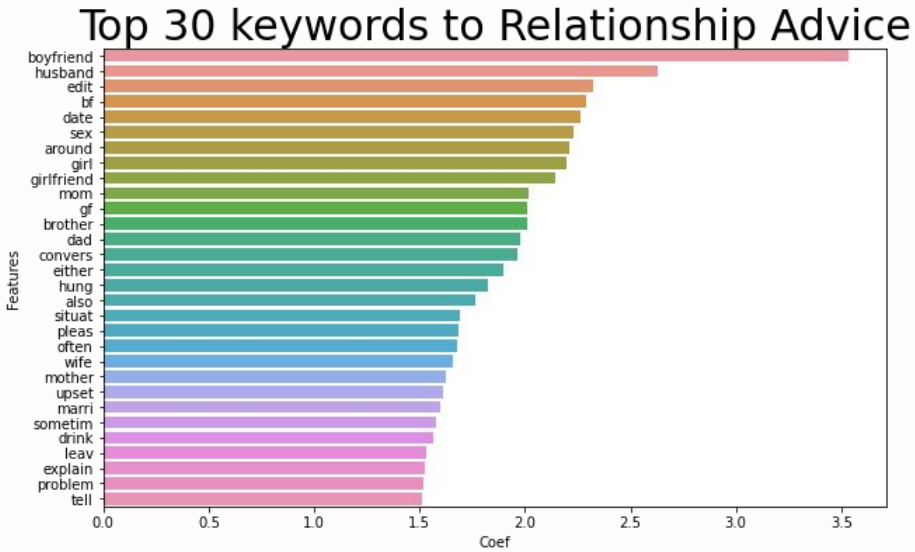
Final Model (TFID Linear Regression)



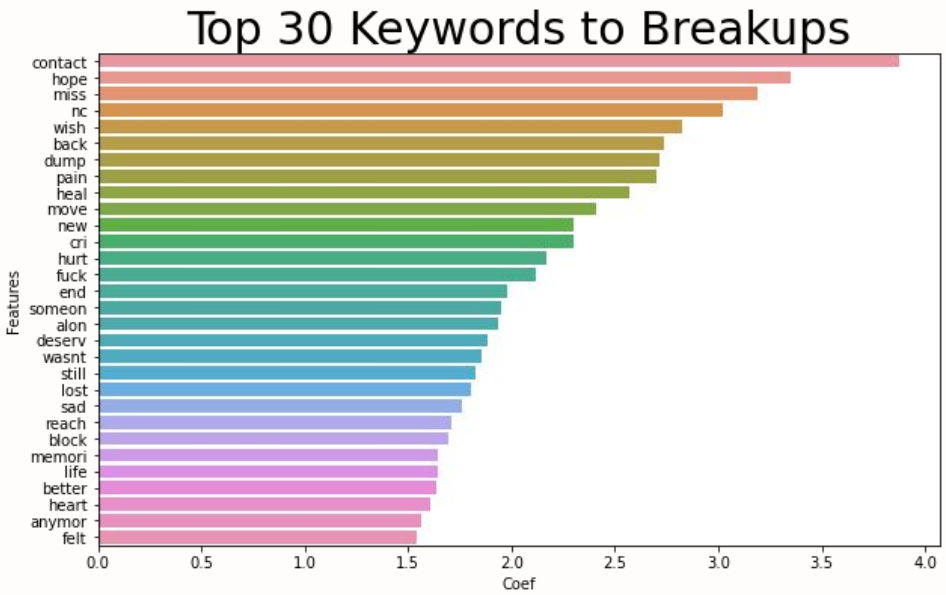


model	TN	FP	FN	TP	T1_acc	T2_acc
LogisticRegression	146	41	26	167	0.865300	0.780000
MultinomialNB	156	31	27	166	0.860100	0.830000

Key findings:
Relationship advice

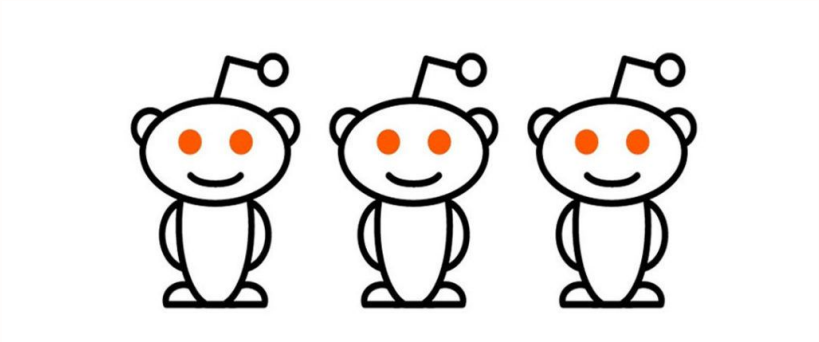


Key findings: Breakup



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
  - Final model