reddit

Reclassification

Juhee Sung-Schenck

TABLE OF CONTENTS

PROBLEM

- Working from home
- Cat jumped on the keyboard
- Categorization removed
- Reclassification needed

WHAT WE CAN DO

Web scraping using pushshift with API

Classification model

PROCESS

01 Collect the data using pushshift API

02 Clean the data

03 Engineer the features and tokenize/lemmatize the texts

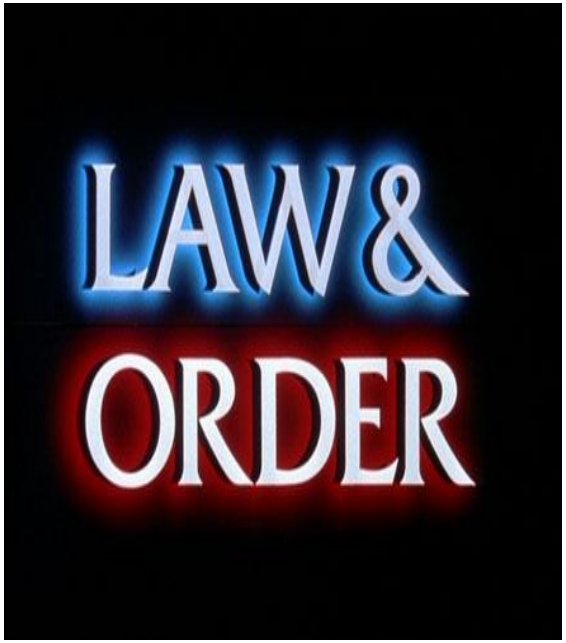
04 Exploratory data analysis

05 Build a few models with hyperparameters

06 Analyze and compare performances

DATA

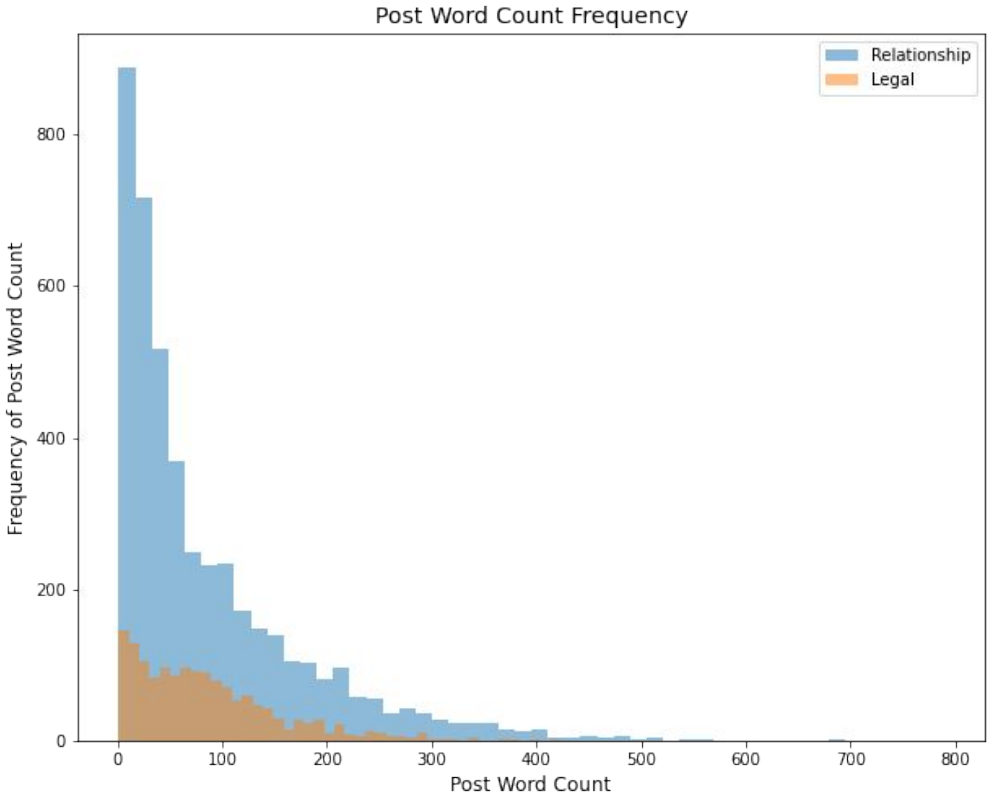
r/legaladvice



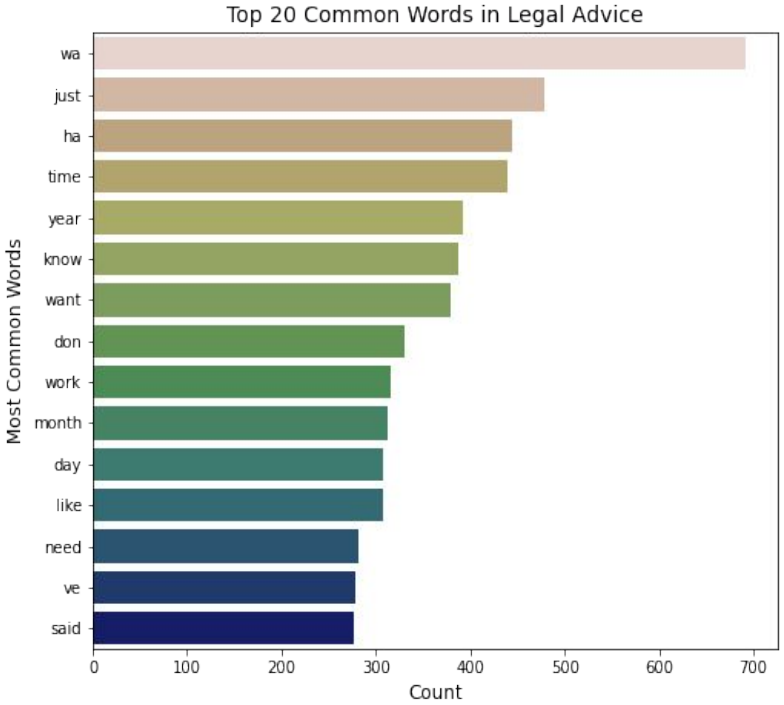
Scraped
title
selftext
body

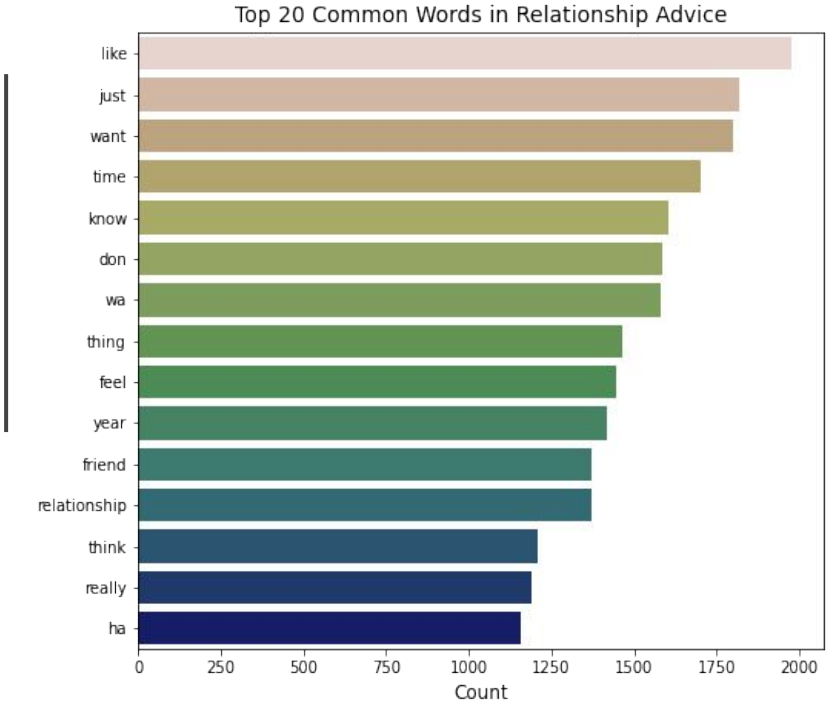
r/relationship_advice

WORD COUNTS

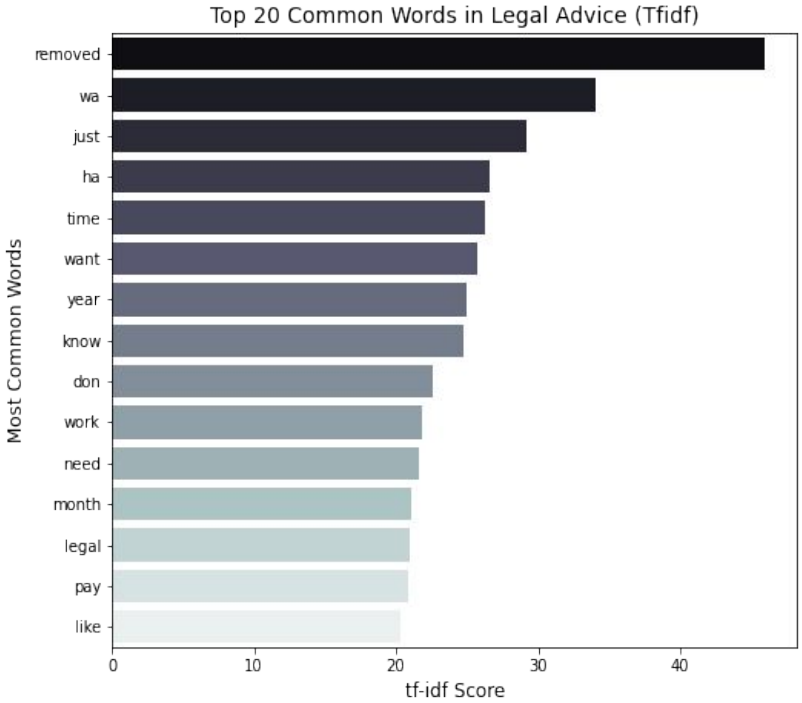

Post Word Count Frequency
Relationship
Legal
Frequency of Post Word Count
0
200
400
600
800
0
100
200
300
400
500
600
700
800
Post Word Count

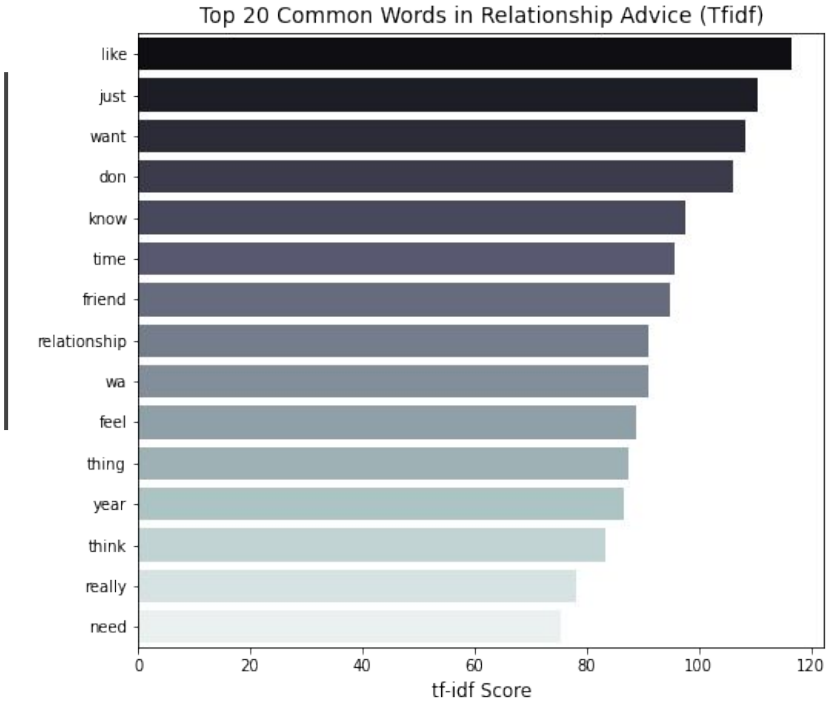
TOP 20 WORDS


Top 20 Common Words in Legal Advice
Most Common Words
wa
just
ha
time
year
know
want
don
work
month
day
like
need
ve
said
0
100
200
300
400
500
600
700
Count


Top 20 Common Words in Relationship Advice
like
just
want
time
know
don
wa
thing
feel
year
friend
relationship
think
really
ha
0
250
500
750
1000
1250
1500
1750
2000
Count

TOP 20 WORDS


Top 20 Common Words in Legal Advice (Tfidf)
Most Common Words
removed
wa
just
ha
time
want
year
know
don
work
need
month
legal
pay
like
0
10
20
30
40
tf-idf Score


Top 20 Common Words in Relationship Advice (Tfidf)
like
just
want
don
know
time
friend
relationship
wa
feel
thing
year
think
really
need
0
20
40
60
80
100
120
tf-idf Score

FINDINGS

Multinomial Naive Bayes

Vectorizer	Count	Tfidf
max_features	6000	1500
min_df	1	0.002
max_df	0.85	0.5
ngram_range	(1, 1)	(1, 1)
stop_words	None	english
score	0.937997	0.91758084

Logistic Regression

Vectorizer	Count	Tfidf
max_features	5500	2500
min_df	2	0.004
max_df	0.85	0.7
ngram_range	(1, 1)	(1, 1)
stop_words	english	None
score	0.917084	0.91633582

FINDINGS

Voting Classifier

Vectorizer	Count	Tfidf
max_features	5500	-
min_df	1	-
max_df	0.8	-
ngram_range	(1, 1)	-
stop_words	english	-
ada__n_estimator	400	
gb__n_estimator	400	-
tree__max_depth	9	-
score	0.90961359	-

Support Vector Machine

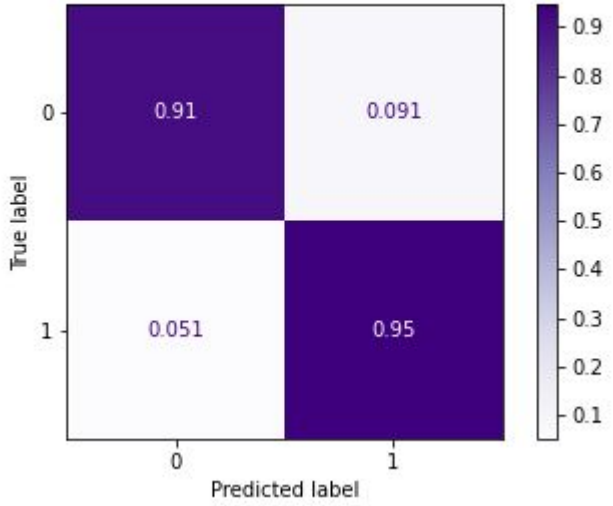
Vectorizer	Count	Tfidf
max_features	6000	5500
min_df	2	0.002
max_df	0.8	0.7
ngram_range	(1, 1)	(1, 1)
stop_words	None	None
svc__C	3.0	2.0
svc__degree	3	3
svc__kernel	rbf	rbf
score	0.90512947	0.92978414

THE BEST MODEL

Model

model	CountVectorizer, Multinomial NB
max_features	6000
min_df	1
max_df	0.85
ngram_range	(1, 1)
stop_words	None
score	0.937997

Confusion Matrix



CONCLUSION

- Use the model that had the best score and the good predictability
- Try this model to different subreddits to improve the sensitivity and specificity

QUESTIONS?