Classification of Subreddits

Project 3
Prerak Agarwal

Problem Statement

Given a post from a user seeking advice on reddit, **is it possible to categorize the post into different subreddits?** If yes, **how accurately** can this advice request categorization be done?

Reddit wants to implement a new “suggestion” feature:

- To guide people seeking advice to appropriate subreddits automatically.
- Advice provided by the dedicated subreddit community would be better and effective.

Approach

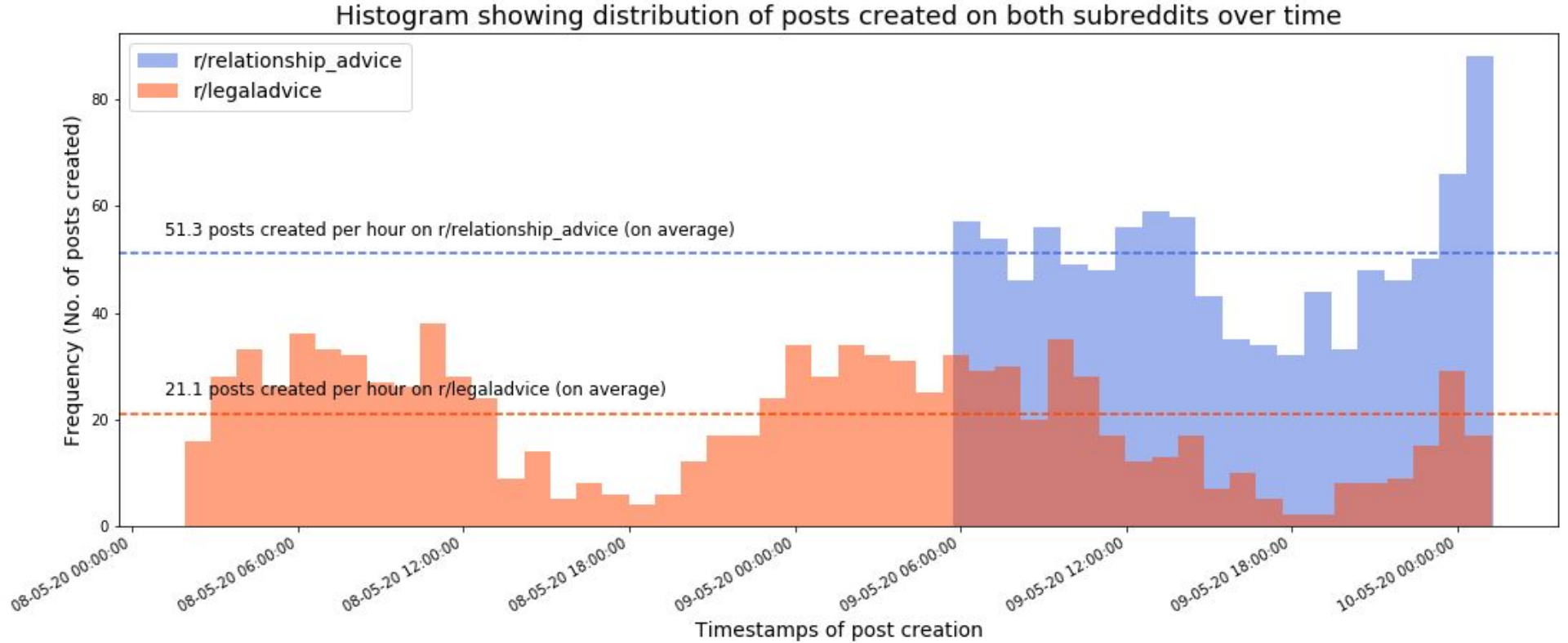
- Chosen subreddits
  - *r/relationship_advice*
  - *r/legaladvice*
- 2000 posts gathered from reddit API
- Posts cleaned and pre-processed
- EDA
- Model prep
- Iterations of classification modelling
- Best model evaluation
- Conclusion

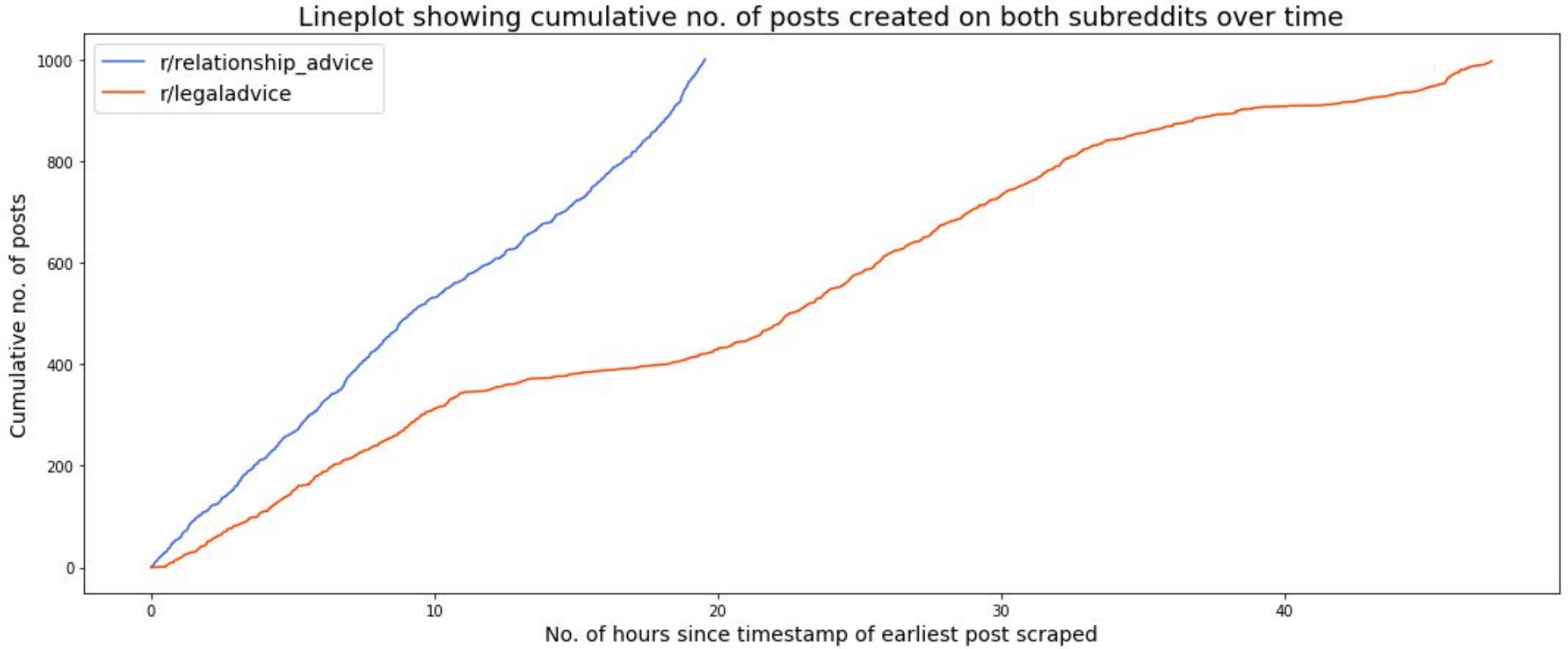
Data Cleaning & Pre-processing

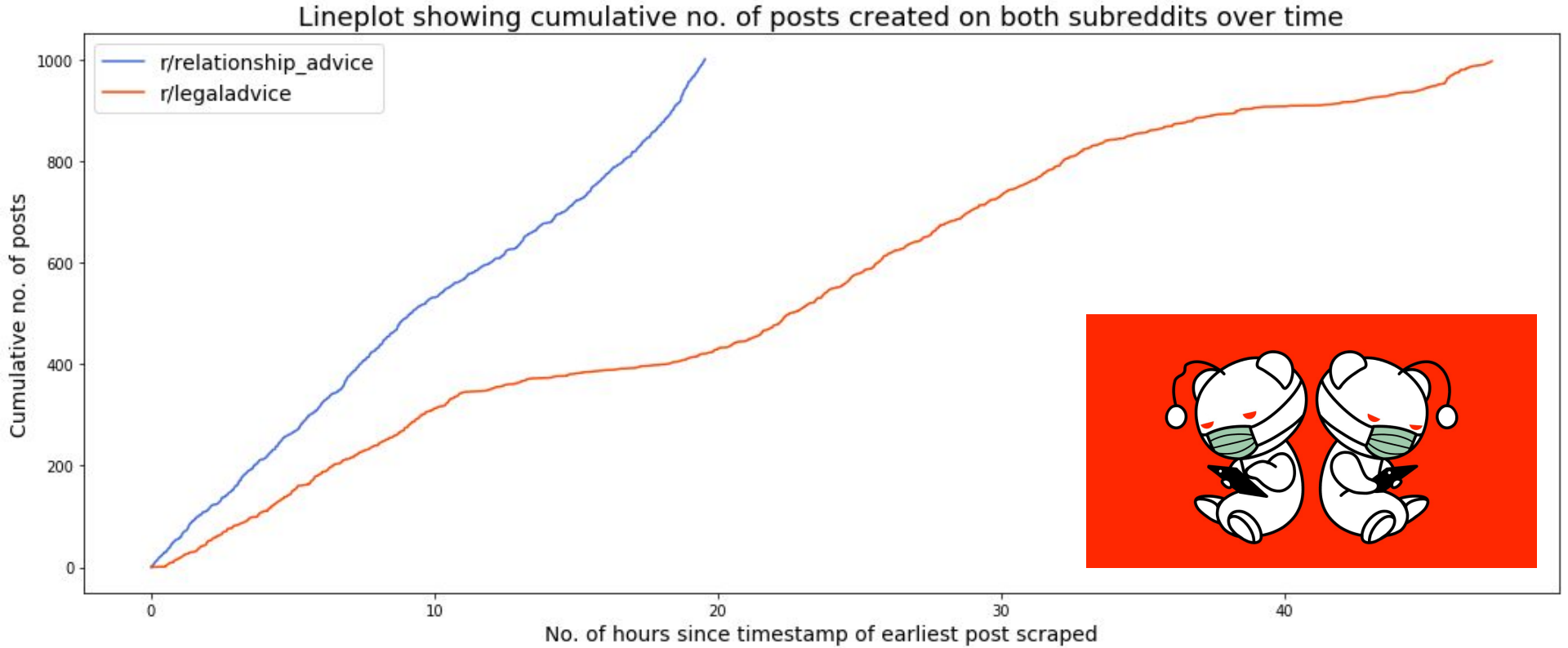
1. Combined post titles and text
2. Removed HTML artefacts (using *BeautifulSoup* library)
3. Expanded all contractions (using *contractions* library)
4. Removed all numbers, punctuations and special characters, except '-' to keep hyphenated words (using *re* library)
5. Converted all text to lowercase
6. Tokenized all words (hyphenated words stay hyphenated) (using *RegexpTokenizer* from *nltk.tokenize*)
7. Removed all stopwords (using the *english* stopwords list from *nltk.corpus*)
8. Removed subreddit names *('relationship', 'legal')* to avoid target leakage
9. Joined all tokenized words into a string separated by spaces
10. After EDA - Stemmed all words (using *PorterStemmer* from *nltk.stem.porter*)

EDA - Background on Subreddits

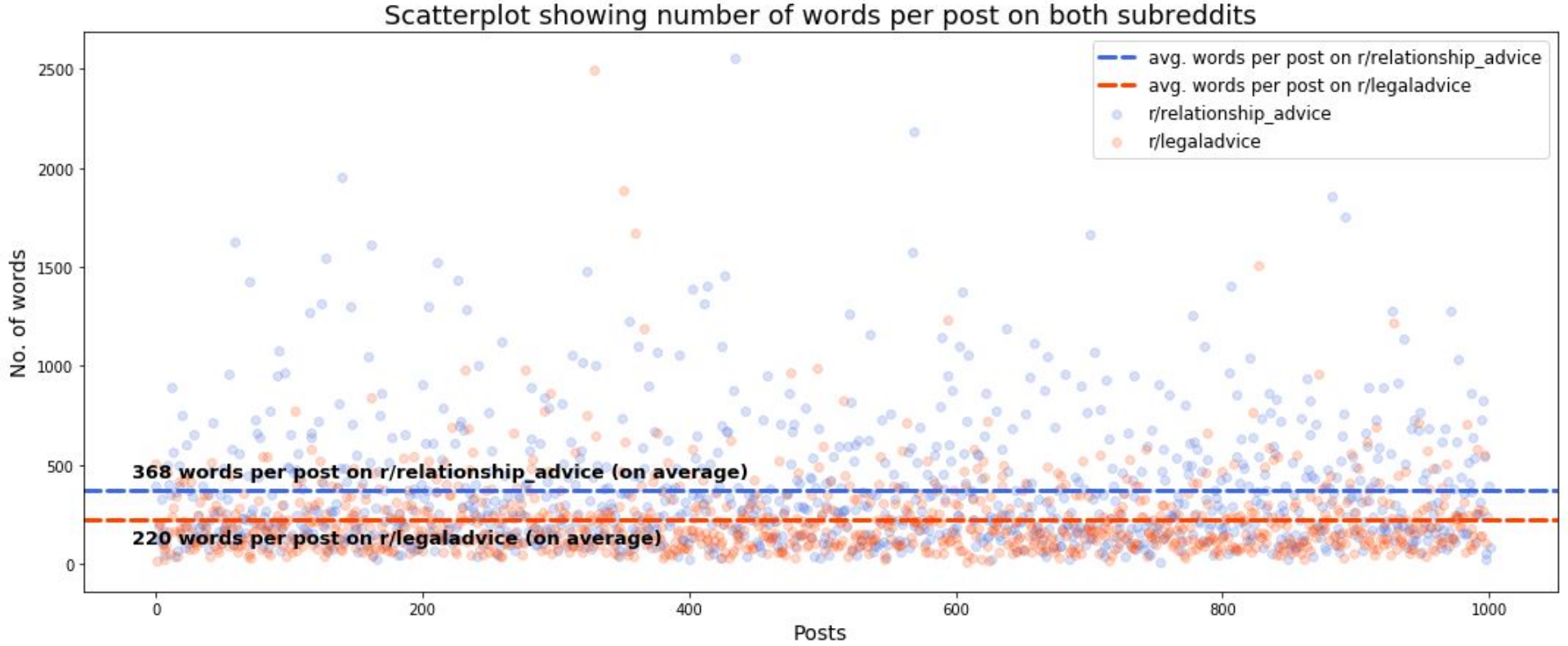
	r/relationship_advice	***r/legaladvice***
Created	14 Jun 2009	26 Oct 2009
Subscribers as of today	> 3 million	> 1.2 million
No. of comments / post (avg.)	9.2	4.5


Histogram showing distribution of posts created on both subreddits over time
r/relationship_advice
r/legaladvice
51.3 posts created per hour on r/relationship_advice (on average)
21.1 posts created per hour on r/legaladvice (on average)
Frequency (No. of posts created)
0
20
40
60
80
08-05-20 00:00:00
08-05-20 06:00:00
08-05-20 12:00:00
08-05-20 18:00:00
09-05-20 00:00:00
09-05-20 06:00:00
09-05-20 12:00:00
09-05-20 18:00:00
10-05-20 00:00:00
Timestamps of post creation

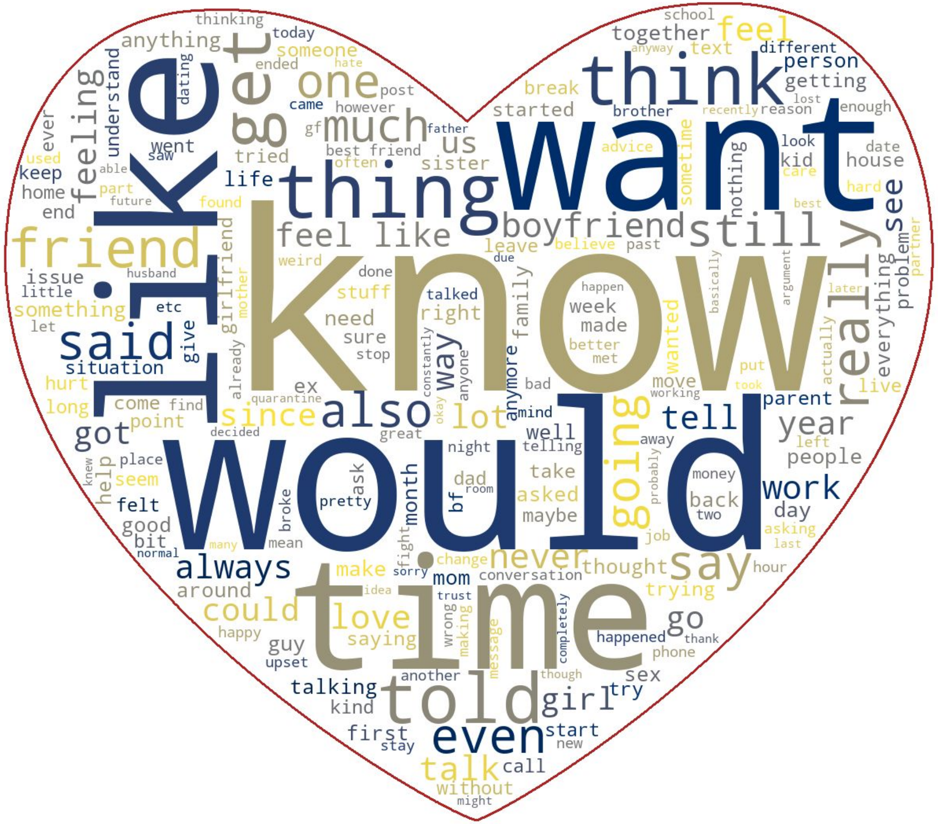

Lineplot showing cumulative no. of posts created on both subreddits over time
r/relationship_advice
r/legaladvice
Cumulative no. of posts
0
200
400
600
800
1000
0
10
20
30
40
No. of hours since timestamp of earliest post scraped

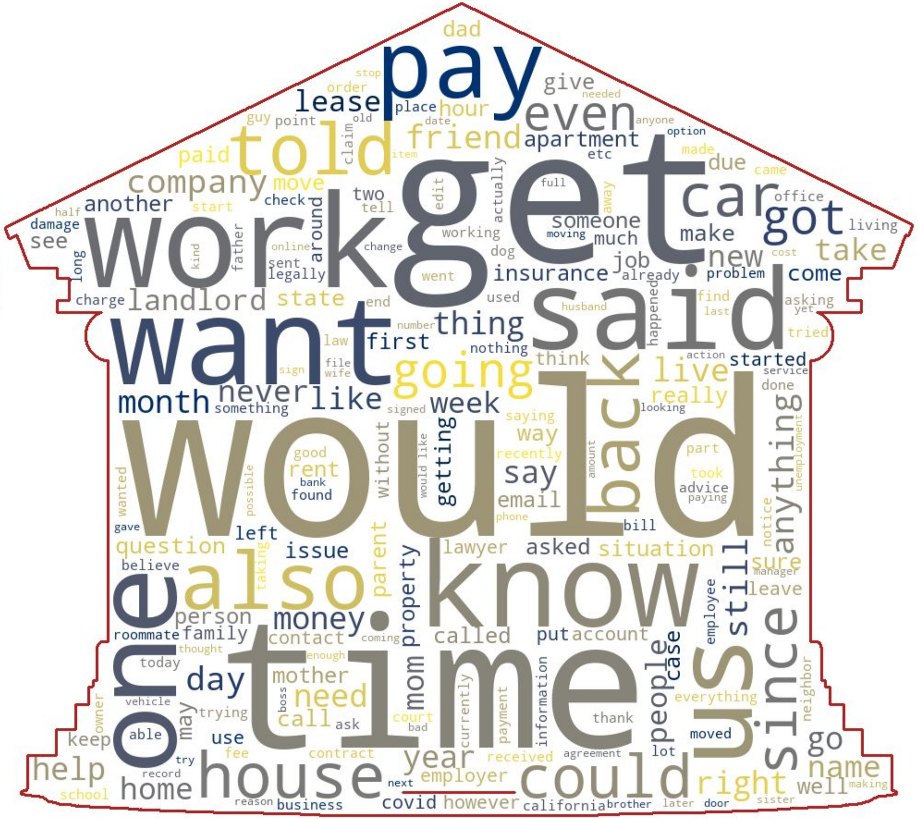


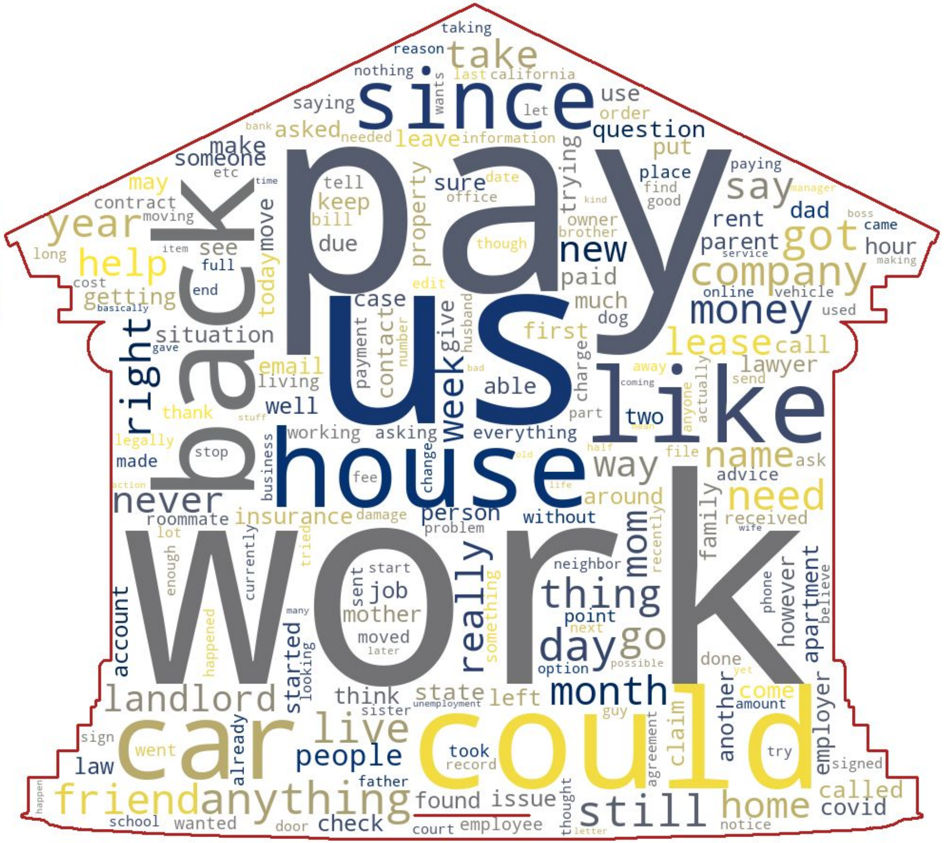
The massive amount of activity on *r/relationship_advice* could be attributed to the fact that more couples are now either *isolated indoors with each other*, or *separated (socially distanced) from each other*, due to the current COVID-19 situation, as suggested in this article.


Scatterplot showing number of words per post on both subreddits
avg. words per post on r/relationship_advice
avg. words per post on r/legaladvice
r/relationship_advice
r/legaladvice
No. of words
Posts
368 words per post on r/relationship_advice (on average)
220 words per post on r/legaladvice (on average)
0
500
1000
1500
2000
2500
0
200
400
600
800
1000

Initial word clouds






like
thing
really
think
friend
still
never
say
boyfriend
feel like
feeling
love
much
work
feel
make
lot
talk
could
always
year
got
go
day
see
people
month
need
tell
girl
asked
sex
guy
anything
family
help
wanted
back
first
since
together
something
way
us
life
thought
happened
understand
talking
keep
home
put
little
everything
done
maybe
stuff
wants
around
seem
long
trying
mom
give
saying
upset
person
dad
happy
pretty
hour
came
break
right
night
kind
found
getting
sure
house
went
ever
great
husband
hurt
sister
parent
better
good
point
bad
live
call
end
felt
best
problem
well
though
leave
without
asking
anymore
change
anyone
advice
month
past
knew
come
look
bit
take
bf
ask
idea
try
every
fight
saw
happy
school
started
sometime
someone
stay
sorry
thinking
mother
issue
partner
new
okay
trust
room
message
making
hard
honestly
weird
made
always
text
care
later
show
job
believe
today
different
working
recently
girlfriend
often
broke
etc
get
looking
nothing
another
conversation
due
dating
week
probably
able
however
though
talked
best friend
decided
money
hate
brother
last
start
basically
two
many
happen
date
situation
telling
let
knows
constantly
phone
ended
gf
used
wrong
enough
mean
move
left
met
future
part
actually
kid
time
either
stop
completely
normal
argument
quarantine
father
seeing
reason
place
already
away
thank
mind
might
use
took
anyway
lost
dad
idea

pay
us
work
could
back
since
house
like
car
friend
anything
landlord
still
home
month
thing
day
go
really
money
company
lease
new
name
need
take
got
say
help
year
right
question
rent
paid
much
way
live
people
week
able
two
around
insurance
never
situation
getting
contract
property
someone
make
asked
keep
bill
due
tell
see
full
end
case
contact
give
first
email
living
well
working
asking
everything
part
lawyer
call
parent
hour
dad
place
find
good
put
use
order
owner
brother
trying
leave
information
needed
saying
taking
reason
nothing
last
california
let
wants
sure
office
though
month
stop
made
legally
thank
fee
change
person
problem
without
received
advice
ask
family
mom
neighbor
job
mother
moved
later
sent
start
point
next
option
done
come
amount
employer
apartment
however
phone
believe
signed
called
covid
notice
employee
court
found
issue
check
record
took
father
sister
think
state
left
guy
claim
agreement
another
try
already
started
looking
account
law
sign
went
school
wanted
door
today
move
business
roommate
lot
enough
currently
happened
tried
recently
file
moving
item
cost
basically
long
etc
may
edit
vehicle
online
service
used
came
making
boss
manager
paying
kind
payment
number
husband
charge
dog
actually
anyone
send
half
mean
coming
away
bad
gave
time
date
bank
many
possible
yet
unemployment
letter
thought
happen
life
damage
action
old
stuff

Model Preparation

1. Combined the two subreddit datasets into one corpus
2. Shuffled rows to mix them together
3. Created target variable
    a. Positive class (1) = *r/relationship_advice*
    b. Negative class (0) = *r/legaladvice*
4. Evaluated baseline accuracy (50.1%)
5. Split corpus into train & test corpora
    a. Stratified according to proportions of classes
    b. Test size of 25%
    c. Training corpus with 1500 documents (posts)
    d. Testing corpus with 500 documents
6. *Accuracy* chosen of evaluation metric

Classification Modelling

Pipelines (Transformer, Estimator)	**GridSearchCV best score**	**Accuracy on training set**	**Accuracy on testing set**
CountVectorizer, LogisticRegression	93.8%	99.8%	93.0%
TfidfVectorizer, LogisticRegression	95.6%	97.9%	96.0%

Classification Modelling

Pipelines (Transformer, Estimator)	**GridSearchCV best score**	**Accuracy on training set**	**Accuracy on testing set**
CountVectorizer, LogisticRegression	93.8%	99.8%	93.0%
TfidfVectorizer, LogisticRegression	95.6%	97.9%	96.0%

Classification Modelling

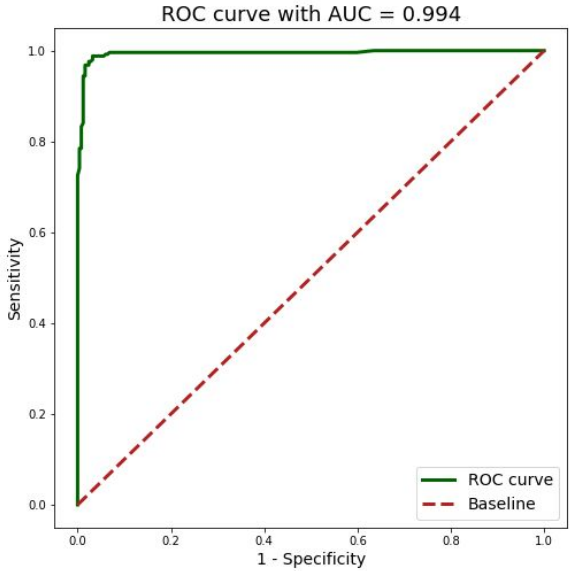
Pipelines (Transformer, Estimator)	**GridSearchCV best score**	**Accuracy on training set**	**Accuracy on testing set**
CountVectorizer, LogisticRegression	93.8%	99.8%	93.0%
TfidfVectorizer, LogisticRegression	95.6%	97.9%	96.0%
TfidfVectorizer, KNeighborsClassifier	91.1%	100%	91.2%
TfidfVectorizer, MultinomialNB	96.9%	97.7%	97.4%
TfidfVectorizer, RandomForestClassifier	92.7%	96.0%	92.0%

Classification Modelling

Pipelines (Transformer, Estimator)	**GridSearchCV best score**	**Accuracy on training set**	**Accuracy on testing set**
CountVectorizer, LogisticRegression	93.8%	99.8%	93.0%
TfidfVectorizer, LogisticRegression	95.6%	97.9%	96.0%
TfidfVectorizer, KNeighborsClassifier	91.1%	100%	91.2%
TfidfVectorizer, MultinomialNB	96.9%	97.7%	97.4%
TfidfVectorizer, RandomForestClassifier	92.7%	96.0%	92.0%

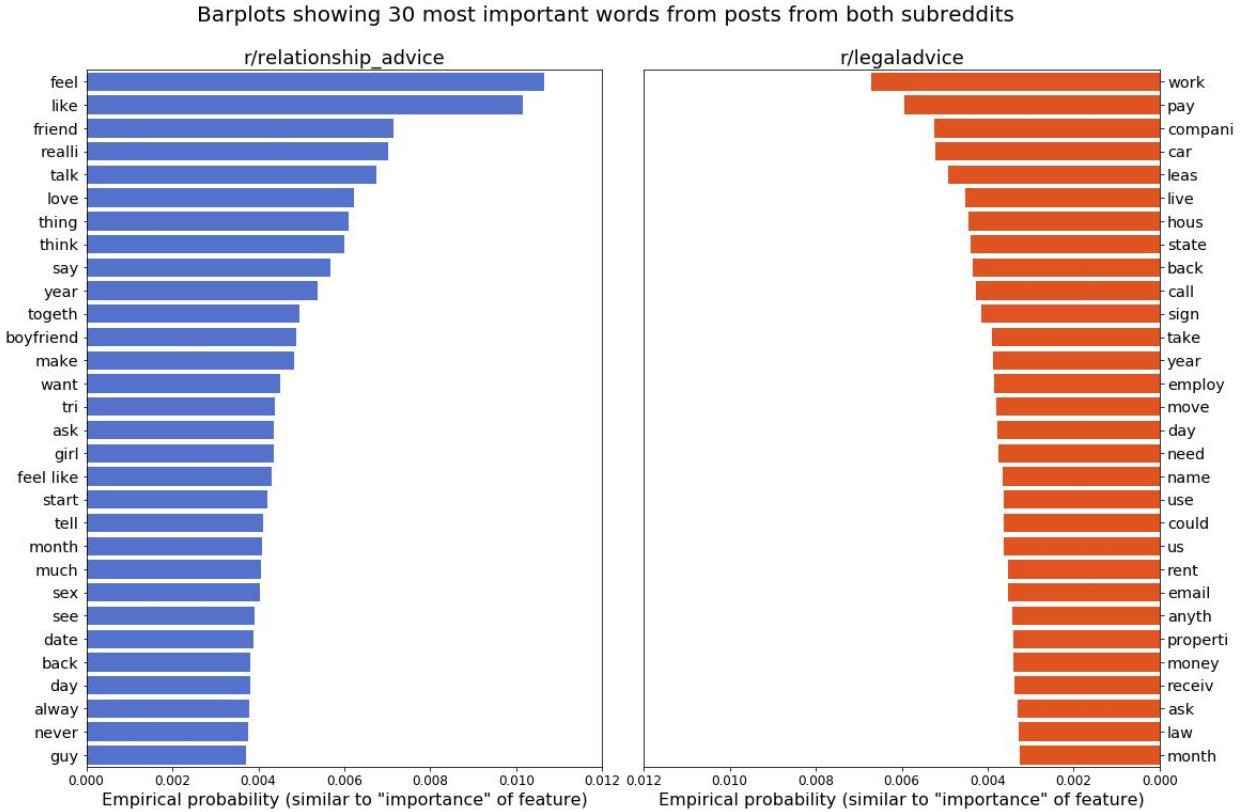
MultinomialNB with TfidfVectorizer

- Training set accuracy = 97.5%
- Est. test set accuracy (cv=5) = 96.6%
- Actual test set accuracy = 97.2%
- Sensitivity (True Positive Rate) = 98.8%
- Specificity (True Negative Rate) = 95.6%
- Total misclassified posts = 14 / 500



Confusion Matrix	Actual r/relationship_advice	Actual r/legal advice
Predicted r/relationship_advice	248 (TP)	11 (FP)
Predicted r/legaladvice	3 (FN)	238 (TN)

Feature Importance



Conclusions & Recommendations

Given a post from a user seeking advice on reddit, **is it possible to categorize the post into different subreddits?** If yes, **how accurately** can this advice request categorization be done?

In case of *r/relationship_advice* and *r/legaladvice*, **classification is indeed possible with a very high accuracy**.

- New “suggestions” feature will be able to guide users to appropriate subreddits very well.
- However, this still needs further study on different combinations of other subreddits.

Thank you!