Natural language processing

PROJECT GOAL

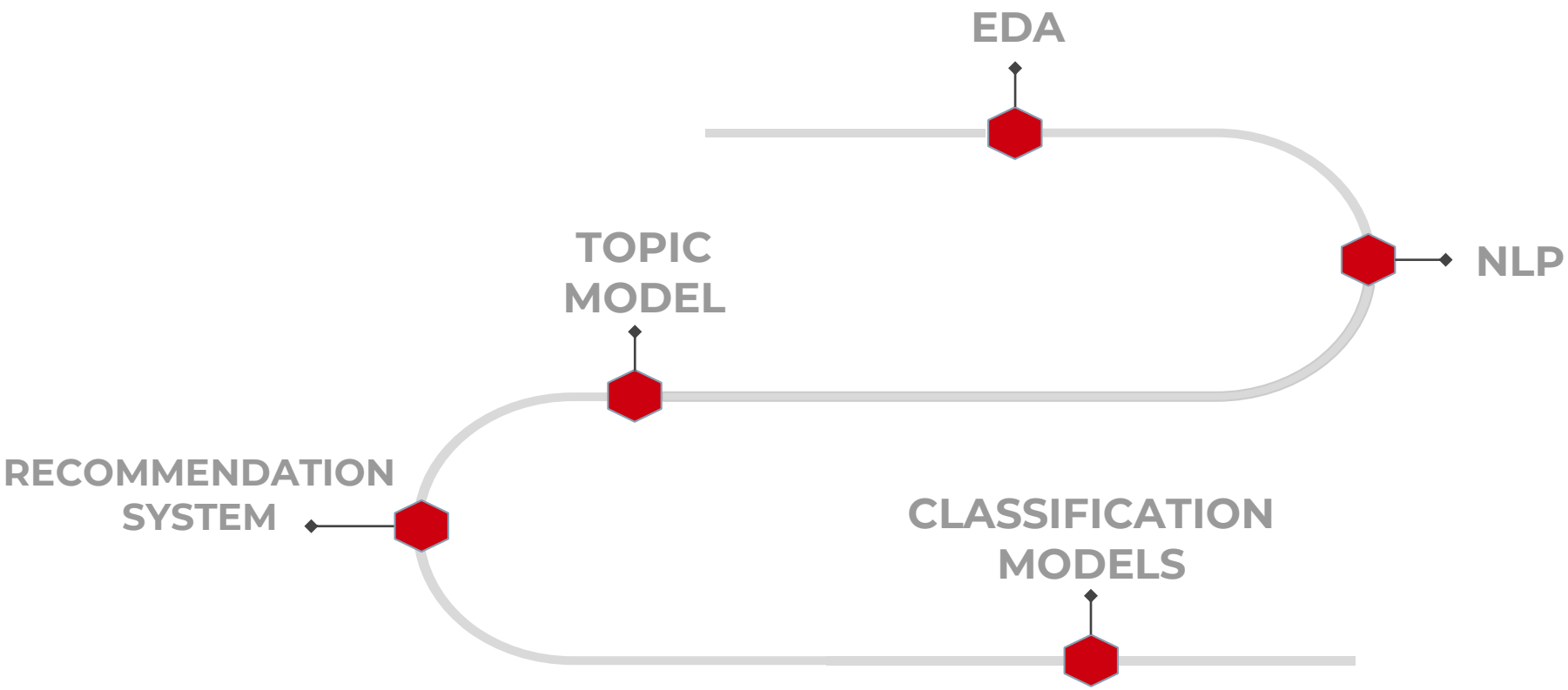
Our goal is to build unsupervised Natural Language Processing (NLP) machine learning models to predict whether a business review text is positive or negative. Also, assigns topics based on the raw text data to find out the business domains and implementing a recommendation system.

DATASET

- Yelp is one of the most famous business review app in the Western Hemisphere countries, with more than 52 million visitors to its mobile sites as of December 2020.

- Two Datasets imported from Yelp website(review & business).

WORK FLOW


EDA
NLP
TOPIC MODEL
RECOMMENDATION SYSTEM
CLASSIFICATION MODELS

WORD CLOUD


food
place
good
one
great
service
restaurant
time
got
go
delicious
table
drink
burger
ordered
order
always
friend
went
come
really
nice
will
came
day
love
well
dont
amazing
back
people
staff
menu
best
try
little
make
made
server
side
im
fresh
want
first
right
perfect
ive
even
bar
dessert
experience
tried
hour
meat
taste
wasnt
debut
visit
flavor
hotel
lot
guy
small
take
told
excellent
another
never
fries
chicken
around
branch
actually
took
everything
inside
fried chicken
asked
milkshake
review
better
thought
dinner
dish
meal
two
much
location
without
party
overall
sauce
night
full
need
say
going
thats
bad
point
worth
hour
lunch
loved
quite
line
favorite
eating
find
still
pretty
cheese
probably
long
wait
though
wonderful
think
wanted
top
sure
give
fun
far
pho
area
price
know
friendly
sweet
thing
minute
spot
way
check
eat
enough
waitress
manager
atmosphere
see
bit
outside
didnt
especially
decided
awesome
reservation
special
kind
something
work
busy
year
half
plate
super
item
stay
home
star
fantastic
beer
left
look
big
now
said
found
put
feel
town
getting
option
everyone
definitely
husband
tell
move
next
seated
looked
water
tasted recommend
first time
problem
extra
music
wish
plus
group
ask
couple
might
coffee
tasty
atlanta
salad
disappointed
usually
pizza
leave
cake
used
maybe
happy
let
broth
door
many
called
parking
three
served
pretty good
different
customer service
happy hour
week
kid
chocolate
helpful
quality
large
customer
chip
sat
walk
name
cool
family
anything
attentive
bartender
front
cold
looking
hot
clean
enjoyed
hostess
fact
someone
crisp
macron
help
hand
charge
ready
local
beef
offer
shake
kitchen
serve
cooked
arrived
youre
couldnt
although
sunday
highly recommend
vancouver
flavorful
will definitely
onion ring
absolutely
appetite
raven
rosebud
dummy
cheap
weekend
started
stuff
seem
thanks
trying
packed
full
curry
hard
bread
able
house
salsa
keep
change
quick
card
close
instead
plus
theyre
dishes
game
business
huge
least
quickly
seemed
coming
waited
breakfast
really good
go back
wife
austin
even though
fast
expensive
okay
pay
regular
sweet potato
literally
either
gave
less
others
n nthe
n nwe
n nmy
n n

TOPIC MODELS

01
LSA

02
NMF

03
CorEx

Best model was Count Vectorizer CorEx with six topics.

TOPIC MODELS

- **(Food):**

sauce,flavor,taste,meat,broth,fresh,crisp,rice,pork,cooked.

- **(NightLife):**

table,seated,bar,restaurant,server,tables,sat,night,waitress,party.

- **(Food_Menu):**

fries,onion,potato,fried,rings,cheese,menu,ots,sweet,milkshake.

- **(Automotive):**

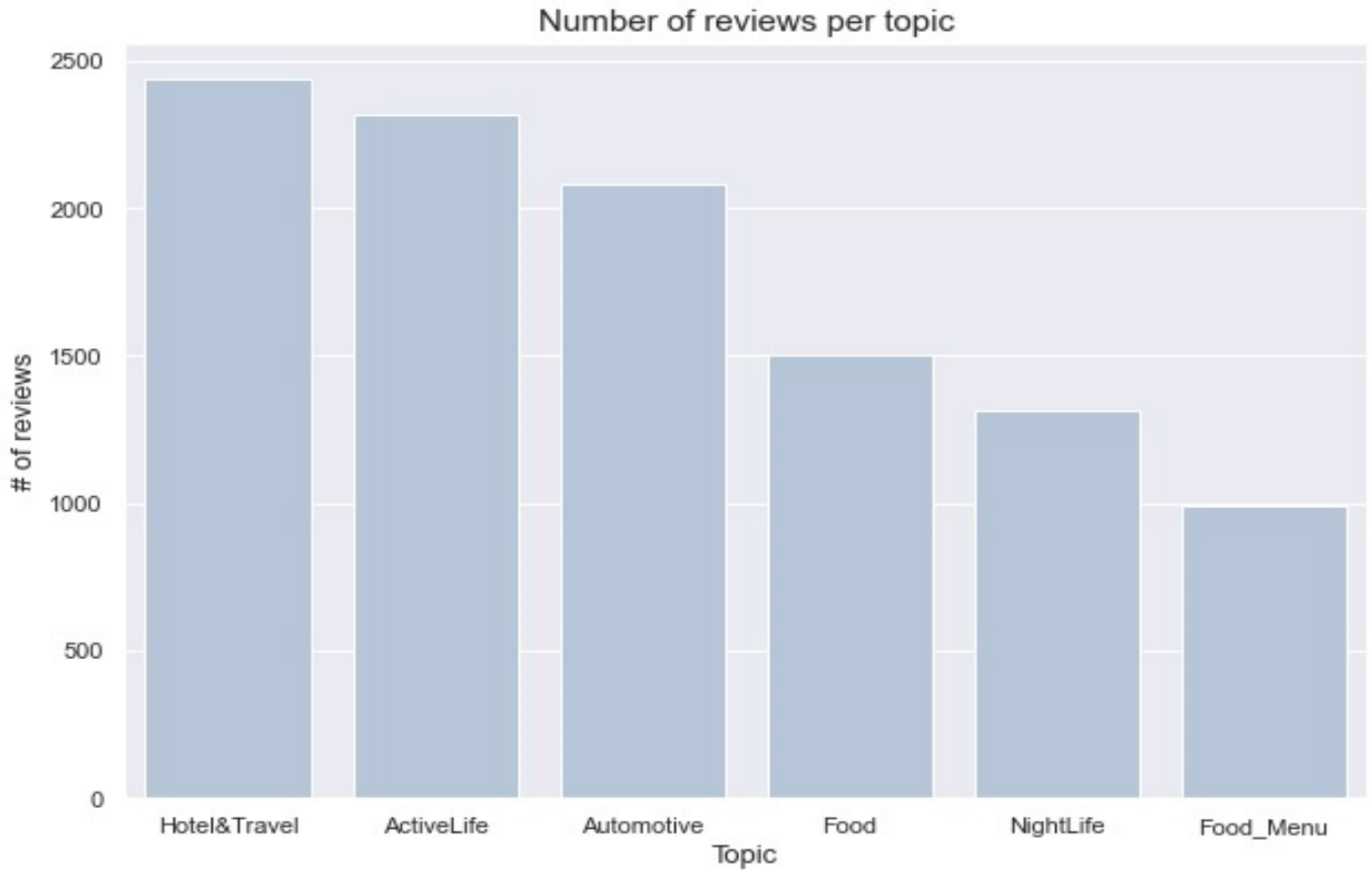
manager,arrived,customer,car,desk,work,rude,phone,guys,received.

- **(Hotel&Travel):**

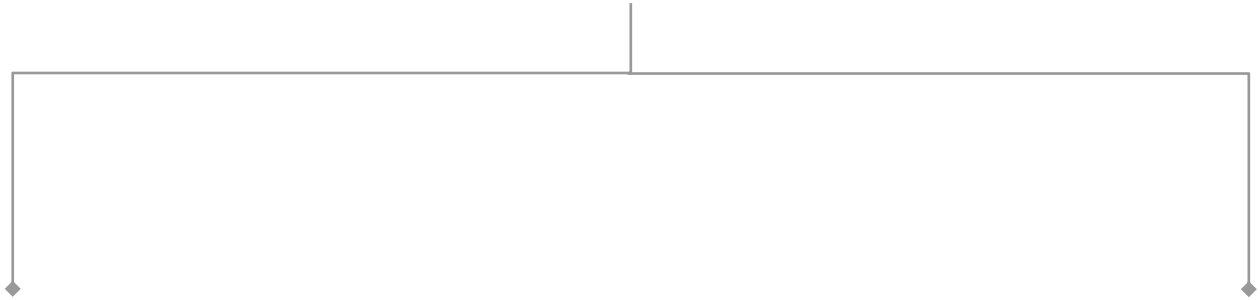
rooms,pool,park,water,stay,resort,area,open,clean,kids.

- **(ActiveLife):**

chicken,music,mac,greens,wales,velvet,busy,bee,irish,red.


Number of reviews per topic
2500
2000
1500
1000
500
0
# of reviews
Hotel&Travel
ActiveLife
Automotive
Food
NightLife
Food_Menu
Topic

RECOMMENDATION SYSTEM



Negative Recommendation System :

- Simple metric

Positive Recommendation System :

- SVD

CLASSIFICATION MODELS

Models	Count Vectorizer		TF-IDF	
	Train	Validation	Train	Validation
Logistic Regression	0.928	0.881	0.913	0.868
Ada Boost	0.880	0.837	0.879	0.843
Weighted Logistic Regression	0.919	0.871	0.912	0.876

TOOLS


jupyter


pandas


NumPy


matplotlib


seaborn


scikit
learn


NLTK


spaCy

CONCLUSION

- Logistic regression was best model.
- For topics CorEx with count vectorizer was the best.
- Yelp reviews is more on Hotel&Travel.

THANKS!