YELP BUSINESS REVIEWS

SENTIMENT ANALYSIS FOR THE PREDICTION OF STAR RATINGS

GROUP G-9

- **Sanket Mhaiskar**
- **Pratish Merchant**
- **Akshay Kamath**

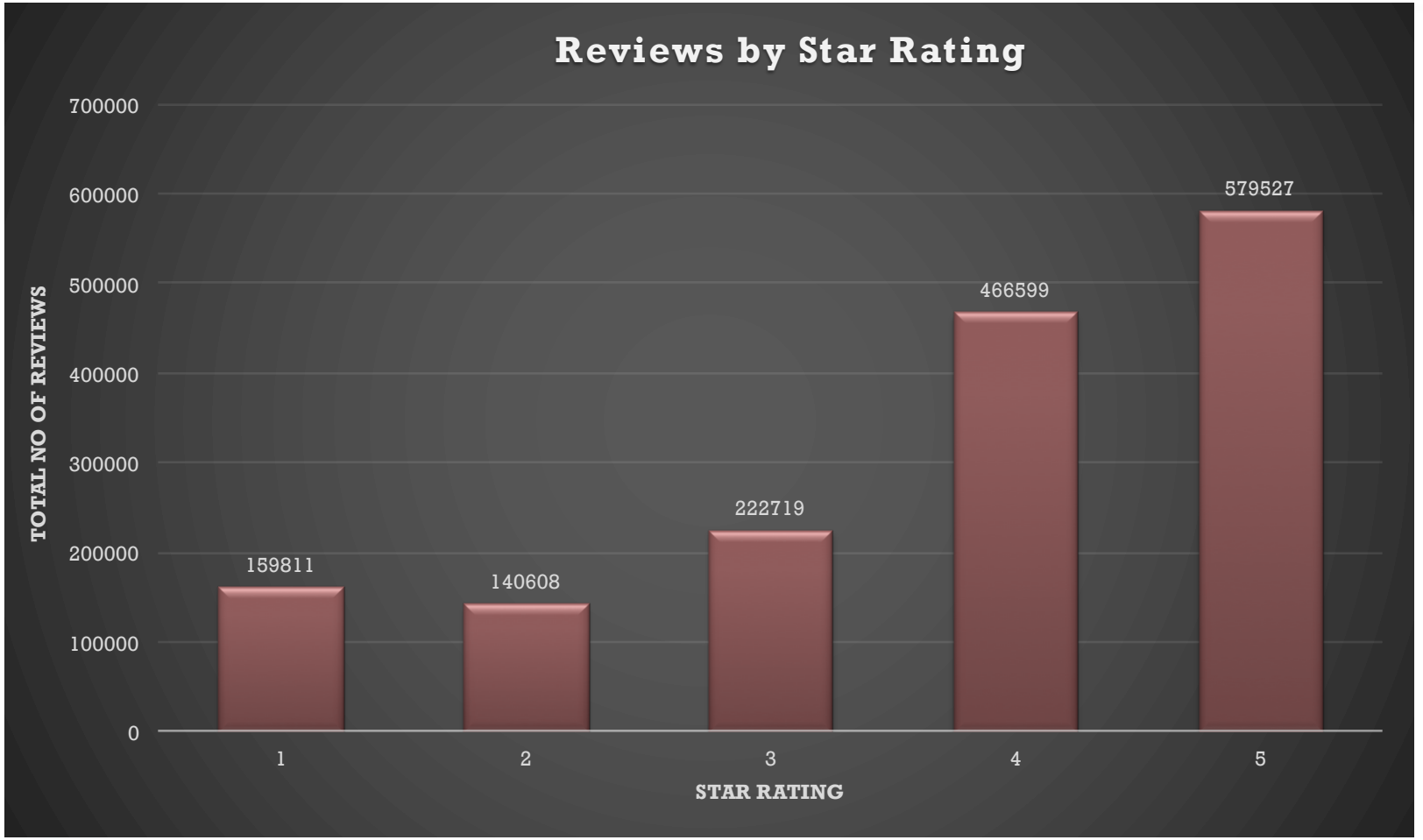
OBJECTIVES

- To identify ratings from reviews.
- Identify the most positive and negative words in the set.
- Perform sentiment analysis task to predict positive or negative emotions.

yelp

DATA

- Yelp data set review as a part of data set challenge.
- Data set consists of the following data fields:
  - Business
  - Check in
  - Reviews
  - Tip
  - Users
- For the scope of this project we mainly focus on user reviews and Business.
- Roughly 25000 reviews are used to perform classification tasks.
- Additionally, we have performed data pre-processing tasks as per our requirements.


Reviews by Star Rating
700000
600000
500000
400000
300000
200000
100000
0
TOTAL NO OF REVIEWS
159811
140608
222719
466599
579527
1
2
3
4
5
STAR RATING

APPROACH

- Initial Approach: Classified reviews into positive and negative and applied classification techniques.

- We have mainly used the following machine learning techniques.
  - Bernoulli's Naive Bayes.
  - Support Vector Machines.

- For each classification technique, we have prepared a data set that maps a review to a sentiment – positive or negative. as follows: Sentiment | Review Text.

- E.g. Positive | We were very pleased with the service!!

SENTIMENT ANALYSIS RESULTS
BERNOULLI NAÏVE BAYES

Accuracy	Majority Classifier
80.10%	79%

	Precision	Recall	F1-Score	Support
Negative	0.50	0.43	0.46	1487
Positive	0.86	0.89	0.88	6014
Avg. / Total	0.79	0.80	0.80	7501

yelp

SENTIMENT ANALYSIS RESULTS SVM

Accuracy	Majority Classifier
90.79%	79%

	Precision	Recall	F1-Score	Support
Negative	0.89	0.61	0.72	1487
Positive	0.91	0.98	0.94	6014
Avg. / Total	0.91	0.91	0.90	7501

yelp

REVIEW PREDICTION

- For Review Prediction, we have mainly used the following machine learning techniques, as before.
  - Bernoulli's Naive Bayes.
  - Support Vector Machines.

- For each classification technique, we have prepared a data set that maps a review to a rating as follows: Review Star | Review Text.
- E.g. 5 | We were very pleased with the service!!

- We then classify each text into bag of words and provide input to the classification algorithms as Unigrams, Bigrams and Trigrams.

REVIEW PREDICTION RESULT
NAÏVE BAYES

Naïve Bayes: Unigrams + Bigrams.
Data Distribution: 80% Training, 20 % Test

Accuracy	Majority Classifier
43.93%	32%

	Precision	Recall	F1-Score	Support
1	0.55	0.32	0.40	706
2	0.28	0.13	0.17	781
3	0.34	0.17	0.22	1343
4	0.43	0.49	0.46	2465
5	0.47	0.70	0.56	2206
Avg. / Total	0.42	0.44	0.41	7501

REVIEW PREDICTION RESULT : SVM

Support Vector Machines: Unigrams + Bigrams.
Data Distribution: 80% Training, 20 % Test

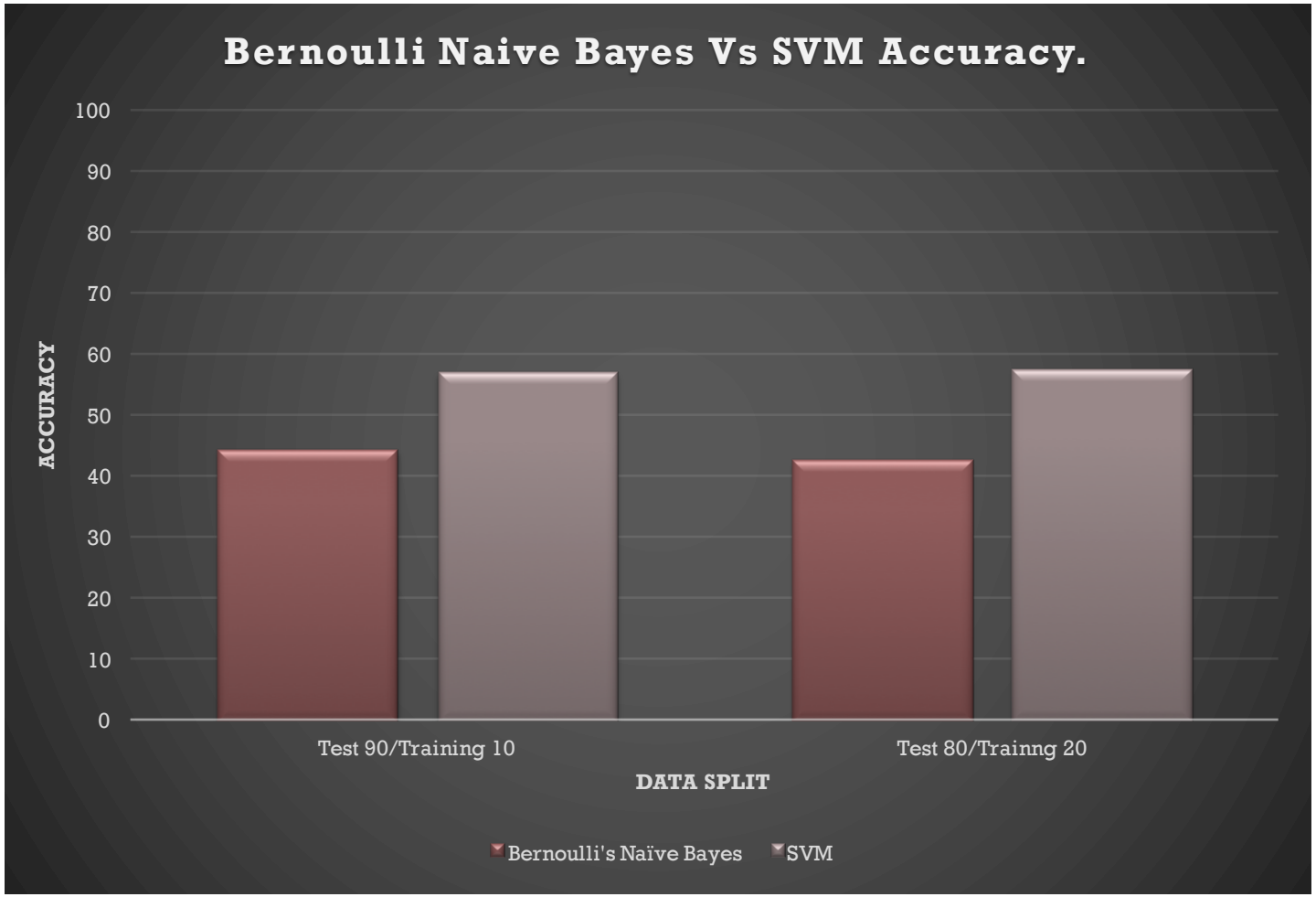
Accuracy	Majority Classifier
56.86%	32%

	Precision	Recall	F1-Score	Support
1	0.75	0.66	0.70	250
2	0.50	0.27	0.35	249
3	0.53	0.35	0.42	469
4	0.49	0.69	0.57	786
5	0.67	0.65	0.66	747
Avg. / Total	0.58	0.57	0.56	2501

N-GRAM RESULT SUMMARY

	Bernoulli NB Accuracy Score	SVM Accuracy Score
Unigram	42.61%	55.31%
Bigram	41.35%	53.57%
Trigram	33.73 %	41.75%
Unigram + Bigram	43.93%	56.85%

- As we can see, we have considered our baseline to be the majority classifier, which comes to 32%
- All classification techniques have performed better than our baseline.
- Using a combination of unigram + bigram has performed better than just unigrams, bigrams or trigrams.


Bernoulli Naive Bayes Vs SVM Accuracy.
100
90
80
70
60
50
40
30
20
10
0
ACCURACY
Test 90/Training 10
Test 80/Trainng 20
DATA SPLIT
Bernoulli's Naïve Bayes
SVM

MOST POSITIVE AND NEGATIVE WORDS

- Used a Dictionary of word – score using Senti-Word list. Scores are between +1 and -1.
- Distribute review into sentences and tokenize into words.Rank words using above dictionary.
- Additionally we can rank entire sentence .
- Output is word , frequency and score.
- Created a word cloud of output using Kumo library.

yelp


struggle
lost
spit
killer
breathe
flea
wit
hate
relaxing
relax
dessert
free
joke
fool
wreck
swear
filthy
thankful
lunchtime
incredible
disappointed
gift
loved
bite
old
kill
bed
steal
aroma
rip
noble
million
fair
rest
drown
not
laughter
kid
whiny
suffer
lazy
fault
food
error
date
trust
dad
fun
noise
lover
prize
soft
sweet
angry
owe
new
honest
divine
dirty
late
value
tree
dick
obese
genius
play
mom
create
cozy
star
good
dance
girl
fairness
suspect
fail
eat
crack
shame
warm
sing
dent
fight
rude
wish
first
save


forced
soon
due
packed
away
another
going
take
back
mean
worse
server
hungry
sorry
unfortunately
wide
buffalo
beat
opened
might
oral
end
left
long
cold
manager
wrong
fire
half
less
hours
wait
hard
price
part
low
veal
still
bad
chain
never
last
bottom
nothing
slow
miss
dirty
final
sandwich
burger
ice
store
sad
thick
stall
fat
expensive
worst
dry
hate
hell
sour
side
pizza
much
like
free
tend
always
burnt
love
first
feel
made
little
tell
hot
red
two
fan
oily
deal
yet
way
bar
lot
let
area
table
better
moved
pool
charge
even
see
good
time
one
know
great
food
place
really
service
ordered
menu
nice
fun
friendly
meal
chicken
restaurant
people
get
order
beer
night
pretty
well
best
new
around
staff
eat
say
waitress
cheese
want
think
find
hour
rock
salad
pastry
fries
drink
sure
day
sing
cant
right
lack
try
overpriced
expected
parking
office
several
average
leave
course
instead
sitting
empty
trying
minimum
front
lots
rather
poor
crowd
problem
terrible
shake
cost
line
rude
bill
shit
old

TOOLS

- Scikit- Learn
- Natural Language Toolkit
- Senti Word Net
- Kumo Word Cloud

yelp

CONCLUSIONS

- We Performed sentiment analysis to classify positive and negative words and got a 10% improvement over baseline using SVM.
- SVM performs better than Naïve Bayes.
- Combination of bi-grams and unigrams gives best results.
- The accuracy is affected with the increase in the number of categories.
- The accuracy increases with increase in data.

QUESTIONS ??????????

yelp

рахмат
danke
謝謝
ngiyabonga
Баярлалаа
spas
teşekkür ederim
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
kiitos
dankie
thank you
gracias
хвала
nanni
nandri
dhanyavad
mauruuru
asante
manana
bedankt
hvala
köszönöm
açiū
obrigada
tenki
bayarlalaa
gracje
dankon
murakoze
enkosi
akun
djiere dieuf
tau
mochchakkeram
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dėkuji
sagolun
taiku
obrigado
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
ありがとう
tanemirt
grazie
didi madloba
najis tuke
мерси
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
mercé
kam sah hamnida
rahmat
ευχαριστώ
xièxie
merci
তোমাকে ধন্যবাদ
감사합니다