F20AA DUBAI UG GROUP 1

PRESENTATION

Table of contents

01

Data Exploration, Visualization And Cleaning

Data exploration

	Score	Review_text
0	5	I received this product early from the seller!...
1	5	***** Numi's Collection Assortment Melang...
2	5	I was very careful not to overcook this pasta,...

(426340, 2)

Get Data

```
DF.isnull().sum()

Score        0
Review_text  0
dtype: int64
```

Check for missing values

Review Rating Distribution

Distribution Analysis

Review Polarity Distribution · Review Word Count Distribution · Review Longest Word Length Distribution

Reducing Skewness

Correlation Matrix

	Score	Word_Count	Longest_Word_Length	Polarity
Score	1	-0.079	-0.092	0.58
Word_Count	-0.079	1	0.38	0.08
Longest_Word_Length	-0.092	0.38	1	-0.022
Polarity	0.58	0.08	-0.022	1

02
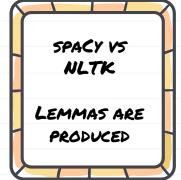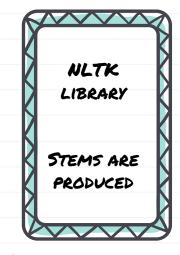
TEXT PROCESSING AND NORMALIZATION

Text Processing and Normalization

spaCy vs NLTK

Lemmas are produced

STEMMING

LEMMATIZATION
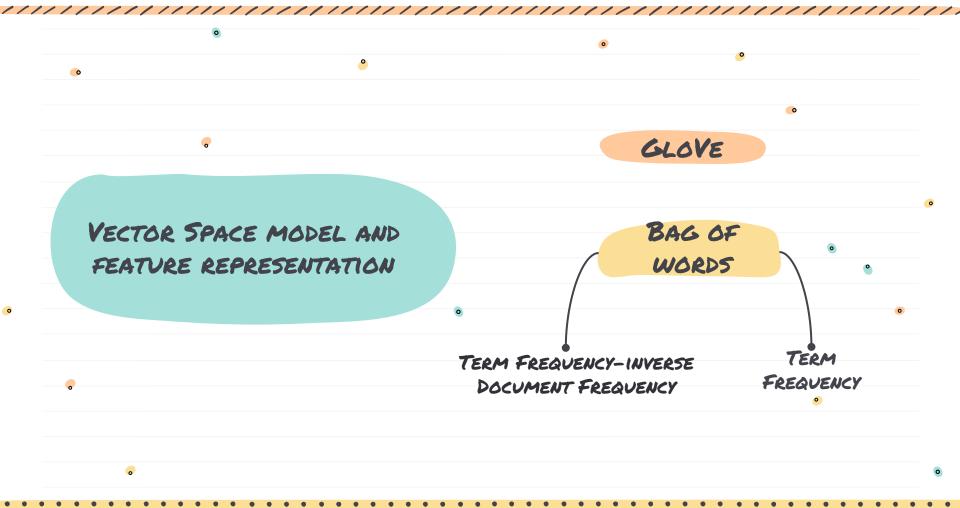
NLTK library

Stems are produced

Lemmatizing process

Stop words removed

Clean reviews

Original reviews

Training Models

03

Vector Space Model and Feature Representation

Vector Space model and feature representation

GloVe

Bag of words

- Term Frequency–inverse Document Frequency
- Term Frequency

GloVe embeddings

```python
from sklearn import preprocessing

def glove_embed(docs):
    ''' Function that converts given documents
        into vectors using a GloVe model.
        @param docs: Documents to embed as an iterable.
        @return docs: Embedded documents as np array.
    '''

    docs_embedded = []
    for doc_words in [doc.split(" ") for doc in docs]:
        doc_vector = np.array([0.0]*100)
        for word in doc_words:
            try: word_vec = GLOVE_MODEL_100[word]
            except: word_vec = np.array([0.0]*100)
            doc_vector += word_vec
        # print(doc_vector.shape)
        # doc_vector = preprocessing.normalize([doc_vector], axis=1)[0]
        docs_embedded.append(doc_vector)
    docs_embedded = preprocessing.normalize(docs_embedded, axis=1)
    return np.array(docs_embedded)
```
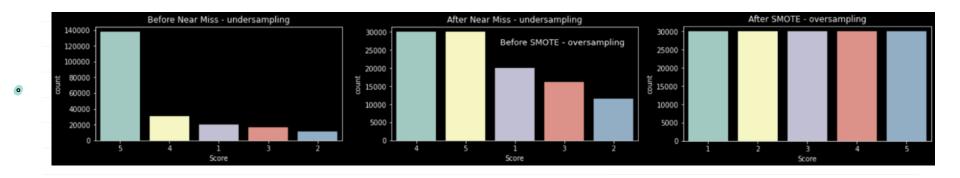
Bag of words Model

TF-IDF

Term Frequency

|	able find	able get	absolutely delicious	absolutely love	actually taste	add flavor	add little	add sugar	add water	almond butter	...	would think	would try	would use	would want	would well	write review	year ago	year old	yr old	yum yum
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Resampling

04
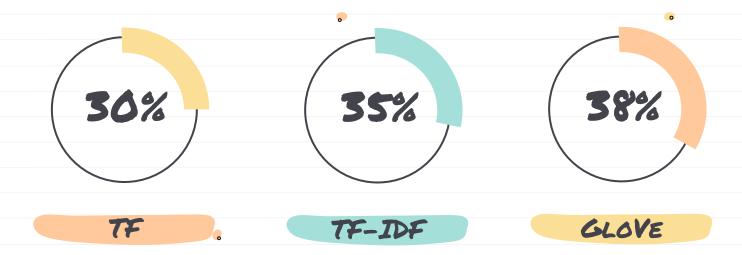
Model Training, Selection and Hyperparameter Tuning and Evaluation
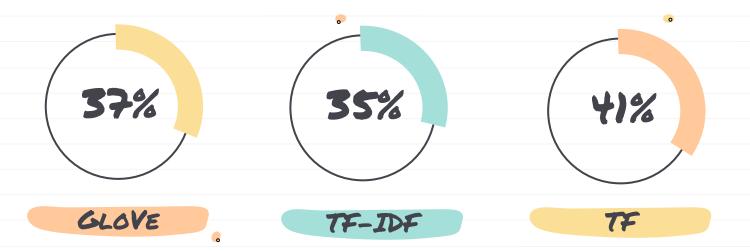
Support vector machines

Gradient descent optimization

40%   50%   36%   55%

Before   After   Before   After

GLOVE   TF-IDF

Logistic regression

30%

TF

35%

TF-IDF

38%

Glove

Naïve Bayes

37%

35%

41%

GLOVE

TF-IDF

TF

```
Accuracy of 10-fold cross validation on testing set:  [0.63183488 0.63053815 0.63756214 0.64015561 0.63849562 0.63665838
 0.6369826  0.63492921 0.6397925  0.64044094]
Mean of 10-fold CV accuracy:  0.6367390025854149
```
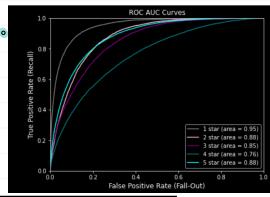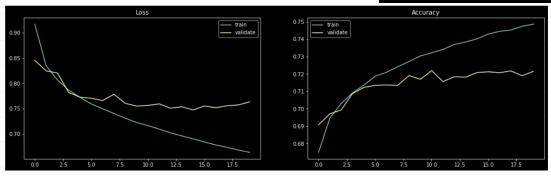
05

Modelling Text as a Sequence

Long-Short Term Memory (LSTM)
Pre-Trained GloVe Embeddings

```
Layer (type)                Output Shape            Param #
=================================================================
Embedding_Layer (Embedding)  (None, 50, 100)        10985400

LSTM_Layer (LSTM)           (None, 100)             80400

Dense_Layer (Dense)         (None, 5)               505

=================================================================
Total params: 11,066,305
Trainable params: 80,905
Non-trainable params: 10,985,400
_____
```
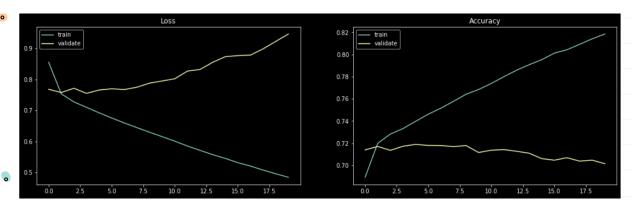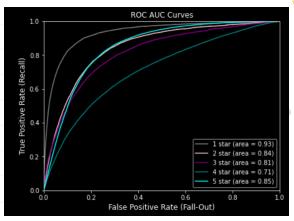
Custom Training of embedding layer

Gated Recurrent Unit (GRU)
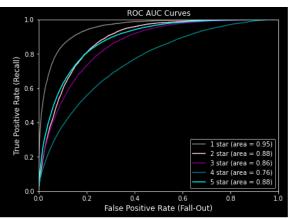
```
Model: "GRU_RNN"
_____
 Layer (type)                Output Shape              Param #
=================================================================
 Embedding_Layer (Embedding)  (None, 50, 100)           10985400

 GRU_Layer (GRU)              (None, 100)               60600

 Dense_Layer (Dense)          (None, 5)                 505

=================================================================
Total params: 11,046,505
Trainable params: 61,105
Non-trainable params: 10,985,400
_____
```
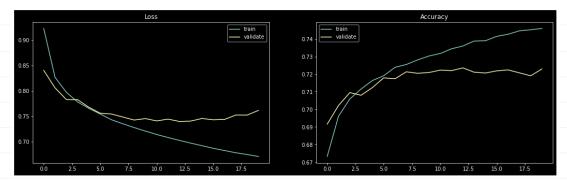
GRU Model

06 Topic Modelling

Topic Modelling of High and Low Ratings

	5 Rated Reviews	1 Rated Reviews
Topic 1	Tea	Eating, Feeding, Hunger, Water
Topic 2	Baked/Baking, Goods, Gluten Free	Meat, Chicken, Leftover, Stale
Topic 3	Pet Food	Mixed, Bad Taste
Topic 4	Soothing, Anti-Allergy, Anti-Acidity	Bad Flavor
Topic 5	Butter, Nuts, Allergy	Digestive Issues, Fruit
Topic 6	Sleep, Relaxation	Coffee
Topic 7	Vegan, Organic	Hexane Additive, Toxic, Pet Food
Topic 8	Female Reproductive Health	Carcinogens
Topic 9	Mixed	Drug Traces
Topic 10	Cheap, Less Nutritional Value, Less Health Value	Stale, Faulty Packaging, Bad Service
Topic 11	Healthy, Snacks	Cost
Topic 12	Pets, Health	Poultry, Ailments
Topic 13	Coconut, Oils, Healthy	Mixed, Pork, Gelatin
Topic 14	Health, Medecine, Desserts	Mixed, Foul Flavor
Topic 15	Mixed, Kids	Foul Flavor
Topic 16	Drinks, Coffee, Herbal Tea	Artificial Sweetness, Bad Taste
Topic 17	Exercise, Diet, Body Weight	Spicy
Topic 18	Pet , Health	Pet Food
Topic 19	Mixed	Pet Food
Topic 20	Pregnancy	Valerian Tea

	5 Rated Reviews	1 Rated Reviews
Broad Topic 1	Tea, Coffee	Eating, Diet, Hunger, Food, Water
Broad Topic 2	Baking/Baked Products, Healthy, Snacks, Special Dietary Needs (Gluten Free, Whole Wheat, etc)	Inorganic Additives, Artificial Sweetness
Broad Topic 3	Pet Food	Stale Food, Meat

20 Topics

Intertopic Distance Map (via multidimensional scaling)

5 rated

Intertopic Distance Map (via multidimensional scaling)

1 rated

3 Topics

Intertopic Distance Map (via multidimensional scaling)

5 Rated

PC2

3

1

PC1

2

gluten free
dog love
dog food
green tea
cat love
cat food
tea bag
cup coffee
peanut butter
earl grey
flavor coffee
black tea
good coffee
love treat
iced tea
green mountain
pill pocket
love tea
great tea
dry food
dog treat
good tea
love food
coconut water
tea great
great coffee
strong coffee
coffee maker
dark roast
love coffee

Marginal topic distribution

2%

Intertopic Distance Map (via multidimensional scaling)

1 Rated

PC2

3

1

PC1

2

cat always
twice day
day cup
eat morning
total eat
spill water
deliberately spill
knock thing
cup total
science diet
eat real
real food
can not
get food
always hungry
always feed
garbage look
wellness indoor
day work
amount science
ingredient science
say cat
food buy
eat habit
please feed
become even
feed little
away eat
choice first
get hungry

Marginal topic distribution

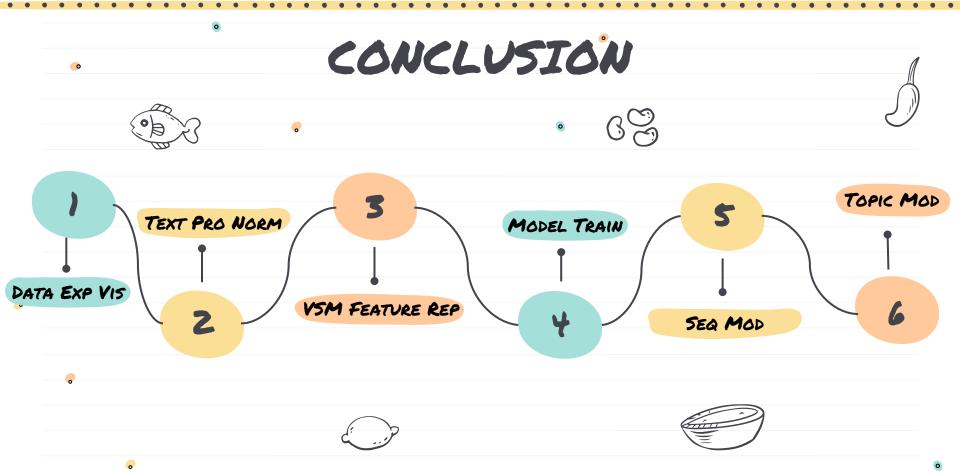2%

5%

TOPIC CONCLUSIONS

Tea
Coffee
Baked
Pet Food
Snacks

Perishable
=> Better
Transport &
Storage

Go
Healthy!

Pet Food!

Junk Vs
Nutritious

Customer
Service
Packaging

Soothing
Medicinal!

Cheap?

CONCLUSION

1 — Data Exp Vis

2 — Text Pro Norm

3 — VSM Feature Rep

4 — Model Train

5 — Seq Mod

6 — Topic Mod

Thanks!

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourcompany.com