Description and Objectives

Exploratory Data Analysis (EDA)

Tools

Data Preparation

Modeling

conclusion

NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl

Description and Objectives

Description

Objectives

The Data & Data source/s

What is our project about?

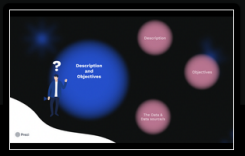
Description and Objectives

Description

Objectives

The Data & Data source/s

- **Get the data**
- **EDA**
  - Data cleaning.
  - Visualization to know the relations between features.
- **Modeling**
  - Data preparation.
  - Apply different classification algorithms.
  - Compare between the results.

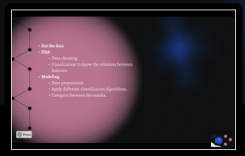


Description and Objectives

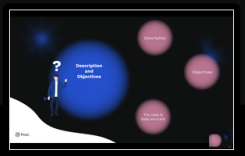
Description

Objectives

The Data & Data source/s



- Source: Kaggle.
- 4 Features .
- 100,000 rows, and each row represents a review.



Description and Objectives

Description

Objectives

The Data & Data source/s

Description and Objectives

Exploratory Data Analysis (EDA)

Tools

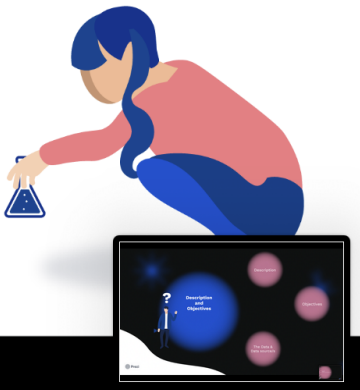
Data Preparation

Modeling

conclusion

NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl



Data Preparation

feature selection

data cleaning

Feature reduction

The final dataset

Feature Selection

Dropping all columns except text and label

Data Preparation

feature selection

data cleaning

Feature reduction

The final dataset

Data Cleaning

- Tokenize.
- Remove punctuations and numbers.
- Remove stop words .
- Convert to lower case.
- Use stemmer
- Create Bag of Words.
- Compute TF.

Data Preparation

feature selection

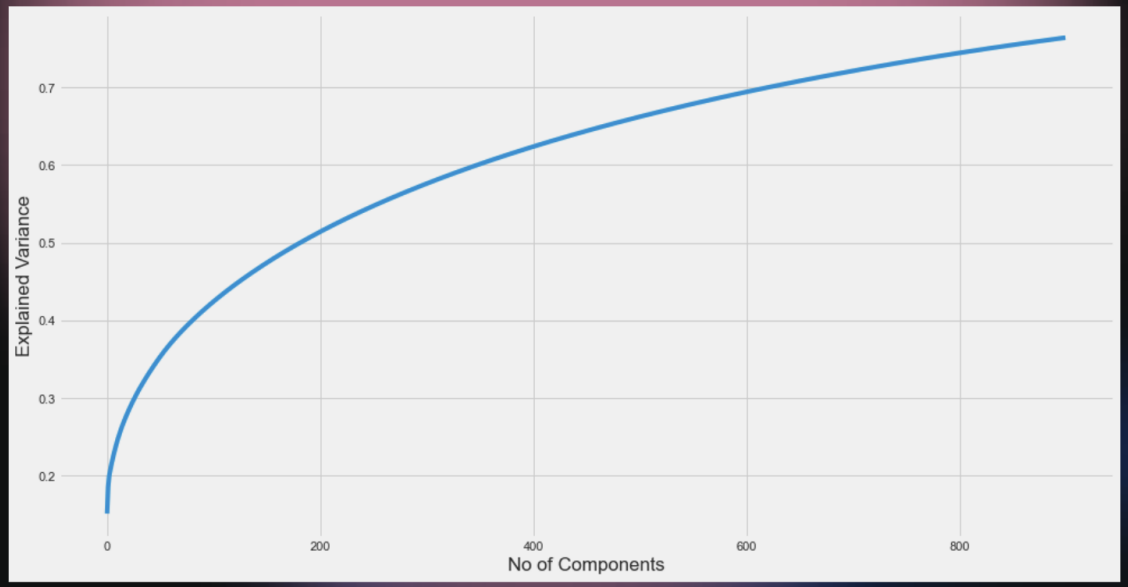
data cleaning

Feature reduction

The final dataset

Feature Reduction

- We applied PCA to reduce the columns to 900.



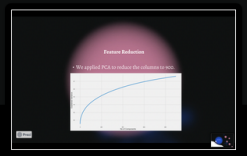
Data Preparation

feature selection

data cleaning

Feature reduction

The final dataset



Final dataset

- 50,000 observation.
- 900 features.

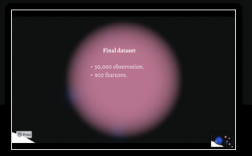
Data Preparation

feature selection

data cleaning

Feature reduction

The final dataset



Description and Objectives

Exploratory Data Analysis (EDA)

Tools

Data Preparation

Modeling

conclusion

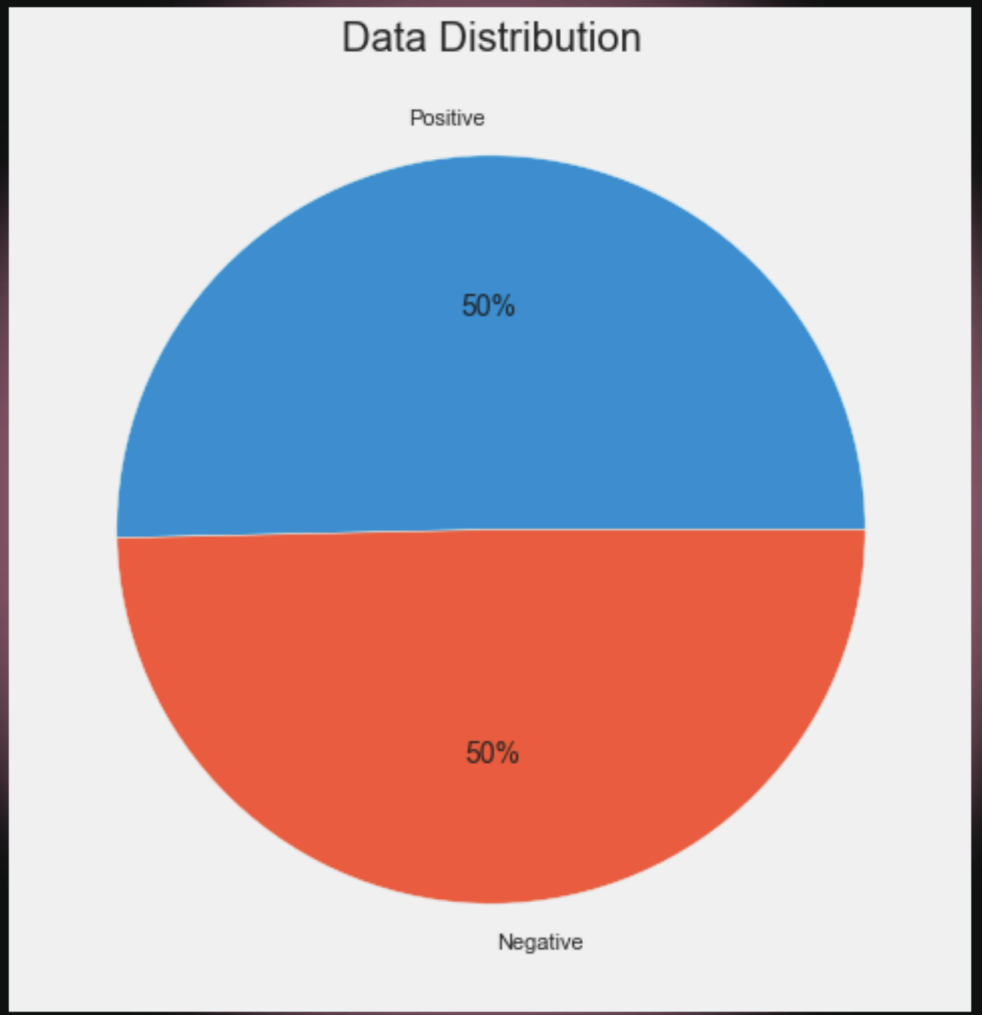
NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl

Exploratory Data Analysis

What is the distribution of data? Is the data balanced or not?

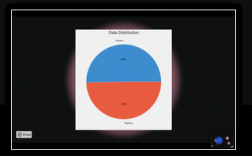
What are the most common positive/ negative adjectives?

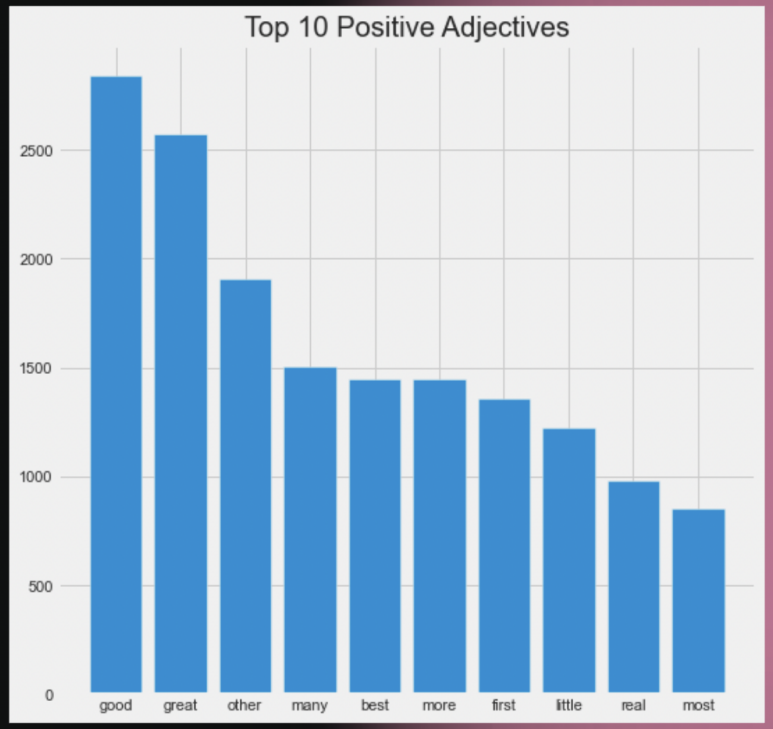

Data Distribution
Positive
50%
50%
Negative

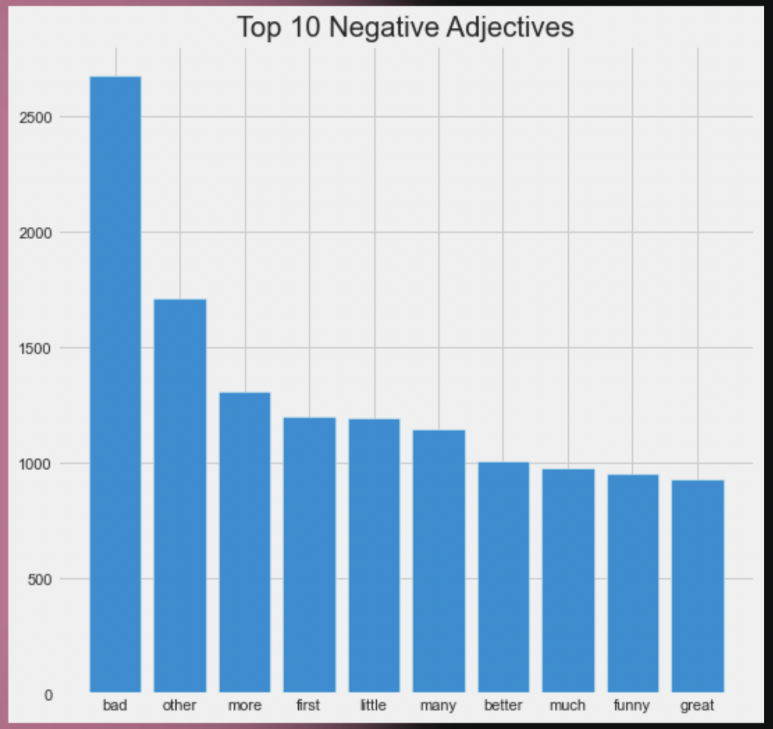
Exploratory Data Analysis

What is the distribution of data? Is the data balanced or not?

What are the most common positive/ negative adjectives?



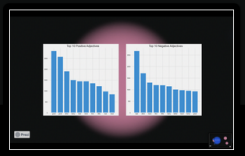

Top 10 Positive Adjectives
2500
2000
1500
1000
500
0
good
great
other
many
best
more
first
little
real
most


Top 10 Negative Adjectives
2500
2000
1500
1000
500
0
bad
other
more
first
little
many
better
much
funny
great

Exploratory Data Analysis

What is the distribution of data? Is the data balanced or not?

What are the most common positive/ negative adjectives?



Description and Objectives

Exploratory Data Analysis (EDA)

Tools

Data Preparation

Modeling

conclusion

NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl

Modeling

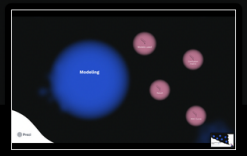
Models used

Confusion matrix

Result

ROC Curve

- KNN
- Logistic Regression
- Decision Tree
- Random Forest



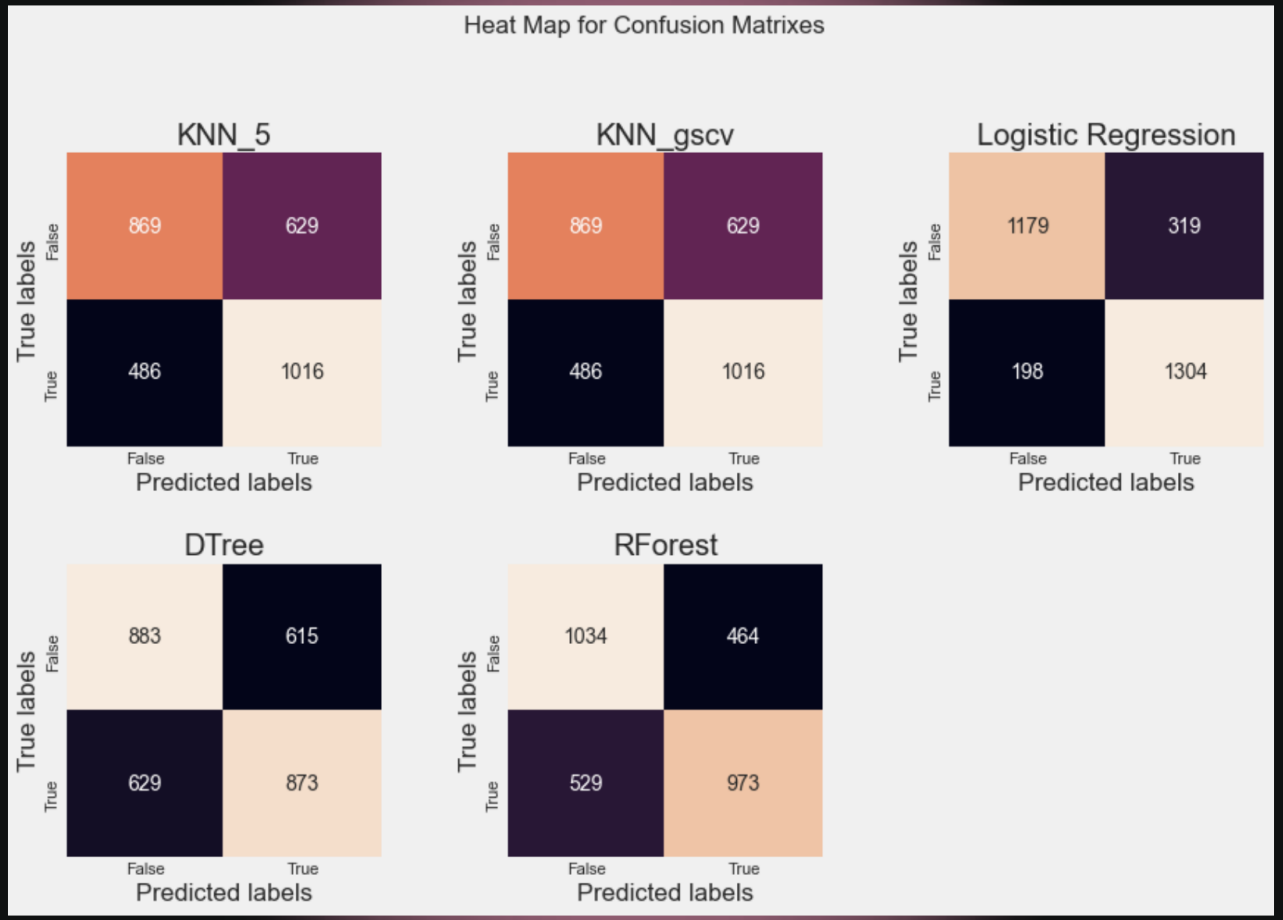
Modeling

Models used

Confusion matrix

Result

ROC Curve


Heat Map for Confusion Matrixes
KNN_5
869
629
486
1016
True labels
False
True
False
True
Predicted labels
KNN_gscv
869
629
486
1016
True labels
False
True
False
True
Predicted labels
Logistic Regression
1179
319
198
1304
True labels
False
True
False
True
Predicted labels
DTree
883
615
629
873
True labels
False
True
False
True
Predicted labels
RForest
1034
464
529
973
True labels
False
True
False
True
Predicted labels

Modeling

Models used

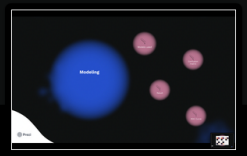
Confusion matrix

Result

ROC Curve

Models Results

Model	Accuracy	Precision	Recall	F1
KNN_5	62.83	61.76	67.64	64.57
KNN_gscv	62.83	61.76	67.64	64.57
Logistic Regression	82.77	80.35	86.82	83.46
DTree	58.53	58.67	58.12	58.39
RForest	66.9	67.71	64.78	66.21



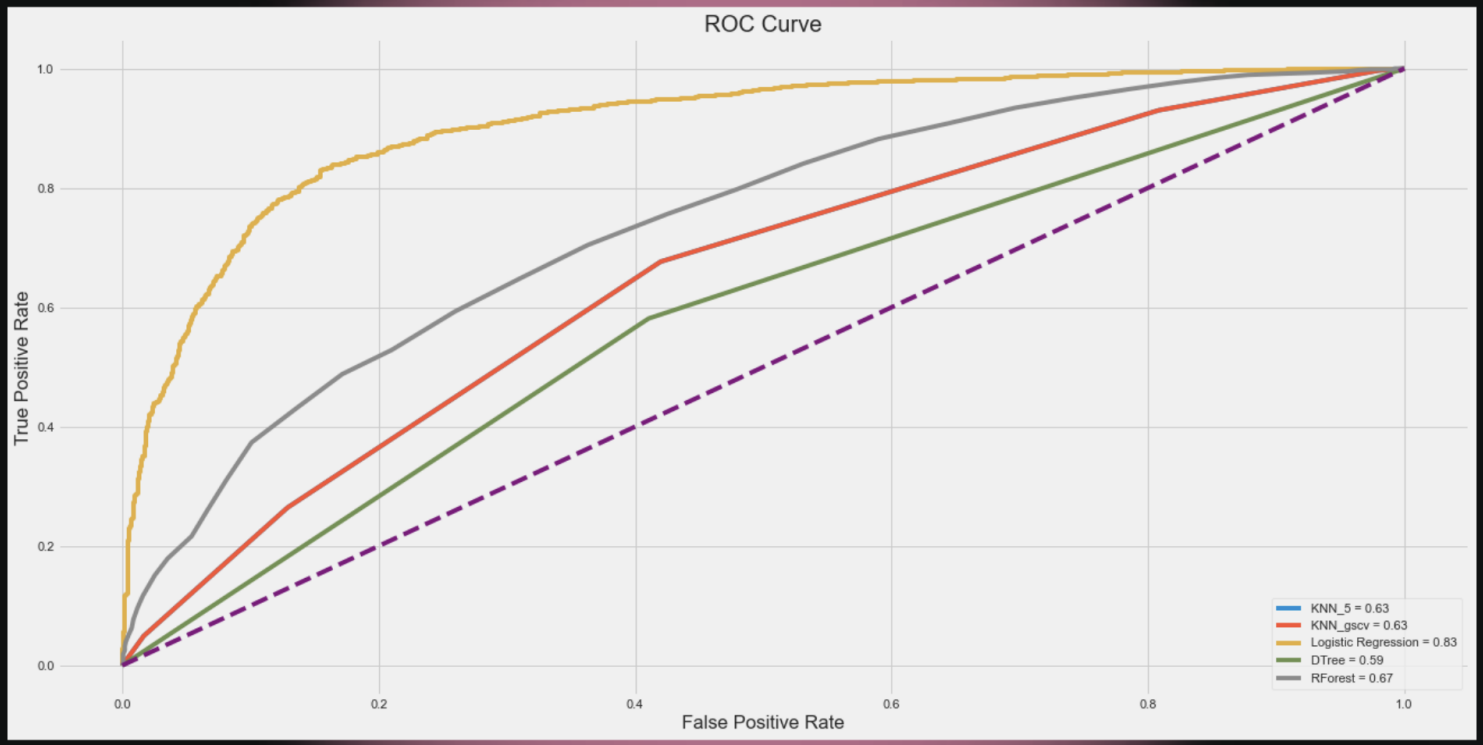
Modeling

Models used

Confusion matrix

Result

ROC Curve


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
KNN_5 = 0.63
KNN_gscv = 0.63
Logistic Regression = 0.83
DTree = 0.59
RForest = 0.67

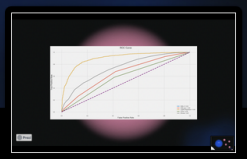
Modeling

Models used

Confusion matrix

Result

ROC Curve



Description and Objectives

Exploratory Data Analysis (EDA)

Tools

Data Preparation

Modeling

conclusion

NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl

Tools

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Sklearn
- SpaCy

Description and Objectives

Exploratory Data Analysis (EDA)

Tools

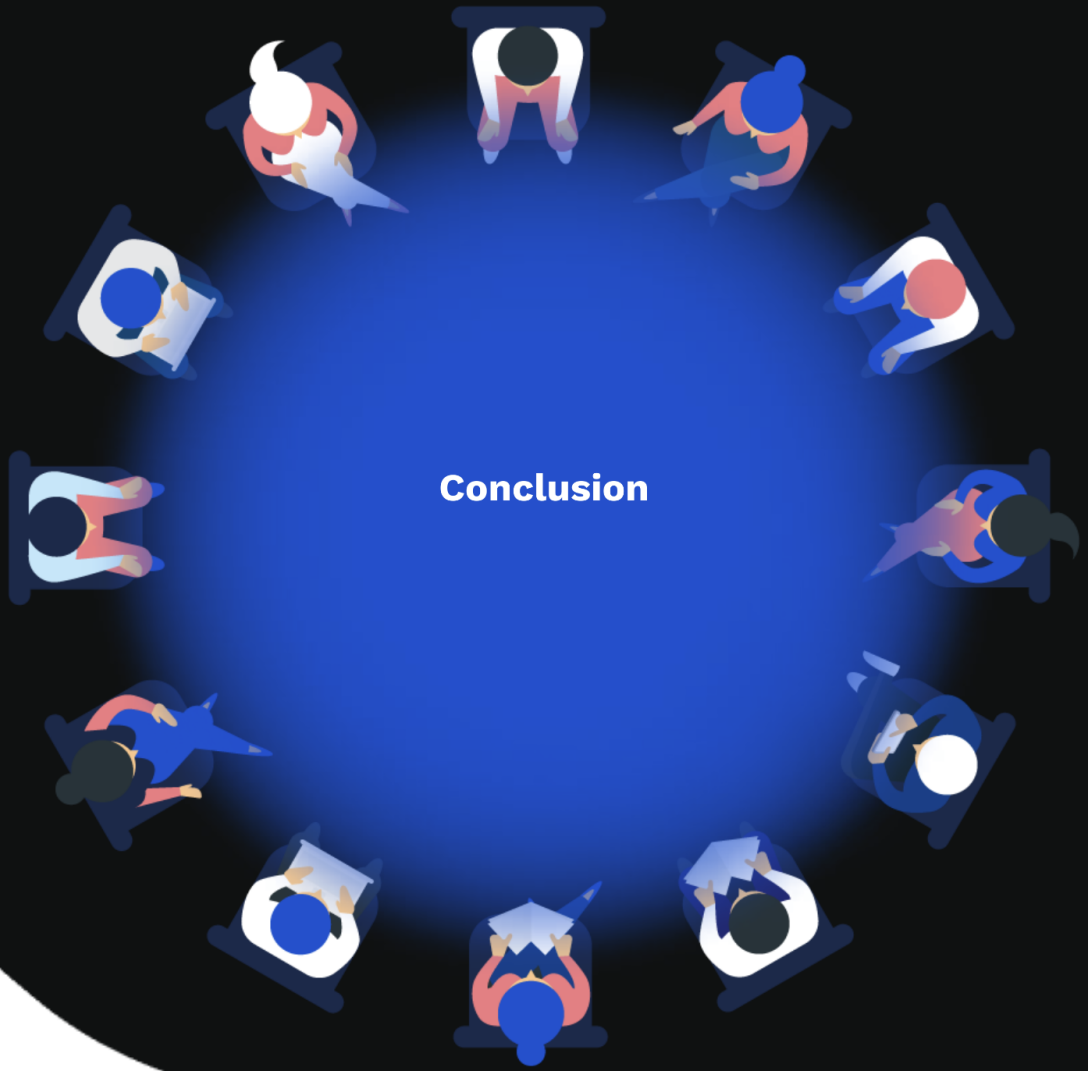
Data Preparation

Modeling

conclusion

NLP Classification for IMDB Reviews

Faisal Alasgah, Ali Altamimi, Saleh Aljomyl



Conclusion