PREDICTING PRODUCT REVIEW SENTIMENT

AYA CHEHABI

SENTIMENT ANALYSIS FOR

- Customer reviews shape purchasing decisions and are abundant on platforms like Amazon.
- Sentiment analysis of product reviews beneficial for prospective customers & businesses:
    - provide valuable insights
    - play a major role in understanding customer satisfaction and preferences.

PROJECT

GOAL: Amazon Product Reviews Sentiment Prediction using Machine Learning

01 Dataset Preprocessing

02 Run Multiple Machine Learning Models

03 Models Performances Comparison

04 Model Deployment

Technologies Used: Streamlit framework along with Python and essential libraries such as NumPy, Pandas, and Matplotlib.

DATA DESCRIPTION

Dataset Kaggle Link for demo
Original Dataset

1597 rows/reviews
27 columns

(product and manufacturer information, **reviews title, reviews text, rating**, metadata, etc.)

1. DATA PREPROCESSING

1. DATA PREPROCESSING

- Filter data to keep relevant columns we want (i.e. reviews, rating, and title)
- Convert all text to lowercase
- Remove URLs and punctuation
- Remove special characters and numbers
- Filter out all non-English words (library: langdetect)
- Remove stopwords (e.g. 'and', 'a', 'the'...)

My interview went well today !! I've got 4 more days to go to know the results & I cannot wait (check the company's website: https://www.clearforme.com)

↓

my interview went well today ive got 4 more days to go to know the results & i cannot wait check the companys website

↓

my interview went well today ive got more days to go to know the results i cannot wait check the companys website

↓

interview went well today ive got days go know results wait check companys website

1. DATA PREPROCESSING

- Filter data to keep relevant columns we want (i.e. reviews, rating, and title)
- Convert all text to lowercase
- Remove URLs and punctuation
- Remove special characters and numbers
- Filter out all non-English words (library: langdetect)
- Remove stopwords (e.g. 'and', 'a', 'the'...)
- Generate word embeddings using Word Encoder or the TF-IDF technique

interview went well today ive got days go know results wait check companys website

Word Encoder (Term Frequency)

- "interview": 1
- "went": 1
- "well": 1
- "today": 1
- "ive": 1
- "got": 1
- "days": 1
- "go": 1
- "know": 1
- "results": 1
- "cant": 1
- "check": 1
- "companys": 1
- "website": 1

1
1
1
1
1
1
1
1
1
1
1
1
1
1

1. DATA PREPROCESSING

- Filter data to keep relevant columns we want (i.e. reviews, rating, and title)
- Convert all text to lowercase
- Remove URLs and punctuation
- Remove special characters and numbers
- Filter out all non-English words (library: langdetect)
- Remove stopwords (e.g. 'and', 'a', 'the'...)
- Generate word embeddings using Word Encoder or the TF-IDF technique

interview went well today ive got days go know results wait check companys website

↓

TF-IDF (Term Frequency - Inverse Document Frequency)

- "interview": 1
- "went": 1
- "well": 1
- "today": 1
- "ive": 1
- "got": 1
- "days": 1
- "go": 1
- "know": 1
- "results": 1
- "cant": 1
- "check": 1
- "companys": 1
- "website": 1

×

- "interview": 1
- "went": 0.5
- "well": 0.5
- "today": 0.1
- "ive": 1
- "got": 0.25
- "days": 0.33
- "go": 0.5
- "know": 0.25
- "results": 0.5
- "cant": 1
- "check": 1
- "companys": 1
- "website": 1

=

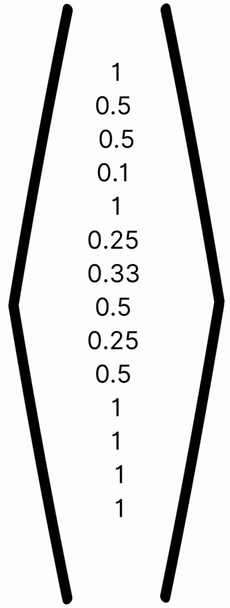
⟨ 1, 0.5, 0.5, 0.1, 1, 0.25, 0.33, 0.5, 0.25, 0.5, 1, 1, 1, 1 ⟩

IDF = log(# sentences/ # sentences containing the word)

1. DATA PREPROCESSING

- Filter data to keep relevant columns we want (i.e. reviews, rating, and title)
- Convert all text to lowercase
- Remove URLs and punctuation
- Remove special characters and numbers
- Filter out all non-English words (library: langdetect)
- Remove stopwords (e.g. 'and', 'a', 'the'...)
- Generate word embeddings using Word Encoder or the TF-IDF technique
- Split the dataset into training and testing sets for model

My interview went well today !! I've got 4 more days to go to know the results & I cannot wait (check the company's website: https://www.clearforme.com)



MACHINE LEARNING MODELS

Sentiment analysis: binary classification (positive/negative)

Logistic Regression

- Models relationship between input feature and binary output
- Estimates the probability of the outcome

Logistic Regression with Stochastic Gradient Descent

- Optimized version
- Updates model parameters iteratively using subset of training dataset

Logistic Regression with SGD & Cross Validation

- Better for model evaluation and hyperparameters tuning
- Splits dataset into training and validation set during multiple iterations

EVALUATION METHODS

Prediction

Ground Truth

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

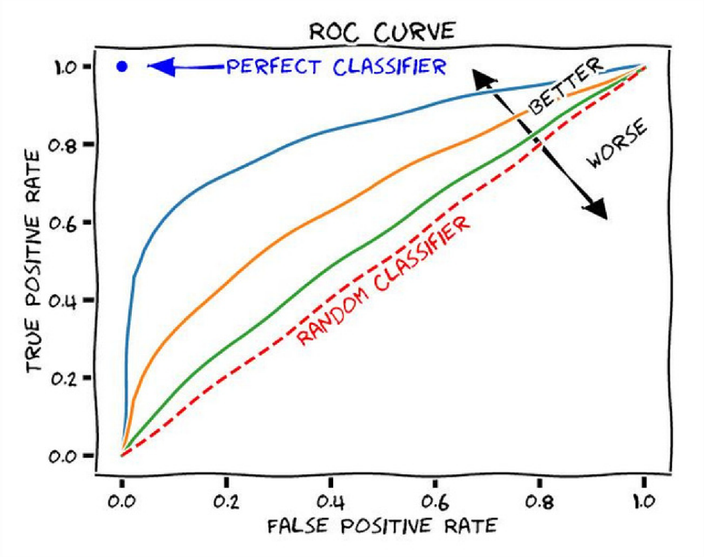
$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Source: https://subscription.packtpub.com/book/big-data-and-business-intelligence/

EVALUATION METHODS

ROC Curve



Source: https://sefiks.com/2020/12/10/a-gentle-introduction-to-roc-curve-and-auc/

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Source: https://subscription.packtpub.com/book/big-data-and-business-intelligence/