Sentiment Analysis on Barbie The Movie Reviews

KenLovers | CDCS2595A | ISP610

Table of contents

01

Introduction

Background, Problems Statements, Objectives,
Literature Review

Introduction

Barbie (2023)

- Director: Greta Gerwig
- Writers: Greta Gerwig & Noah Baumbach
- Main Casts: Margot Robbie, Ryan Gosling, Issa Rae
- Genres: Adventure, Comedy, Fantasy
- Overall Review: 6.9 / 10 (IMDb)
- Awards: 106 wins & 325 nominations

'She's everything. He's just Ken'

Sentiment Analysis

Definition

Also known as opinion mining, a natural language processing techniques that involves determine the sentiment or emotional tone expressed in a piece of text

Application

Widely used in various industries and applications such as business for feedback, social media to understand public opinion on different topics

Challenges

Complexity of language such as sarcasm, irony and ambiguity. Sentiment can vary based on individual perspectives.

Techniques & Model

Rule-based systems to machine learning and deep learning approaches. Rule-based rely on predefined rules and dictionaries while ML trained on labeled dataset.

Problem Statement

Ambiguity and Contextual

- Face difficulties with uncertain language, sarcasm and contextual complexities particularly in movie reviews.
- Need of model that effectively understand and interprets the intricate aspects of language in movie critiques.

Performance of the models

- Some of the models having difficulties to classify the result accurately as it may struggle to understand the language nuance.
- Performance evaluation of the models must be done to determine the best model to use for sentiment analysis for movie critiques.

Objectives

Model Understanding

To develop and train a model for movie reviews that uses advanced techniques to analyze sentiment and comprehend complex language.

Sentiment Classification

To determine whether the majority of reviews express positive or negative sentiments toward Barbie movies.

Evaluate and Compare

To evaluate and compare different models of sentiment analysis for movie reviews using accuracy, recall and precision.

Literature Review

Literature Review	Topic	Citation
Research has determined that reviews that present both sides of an argument are more beneficial to users than reviews that contain a bias.	Domain	(Kong et al., 2023)
Data obtained from a representative set of online reviews penned by IMDb users residing in the United States, as well as Douban users based in China, were consolidated.	Dataset	(Lou et al., 2023)
Users utilize platforms such as social networks, review websites, forums, and blogs to express their views on products and services. Analyzing user expressions can provide valuable patterns for decision making.	Dataset	(Saraswathi et al., 2023)
The Sentiment Analysis technique extracts relevant information from User Reviews dataset and classifies them into positive and negative comments for decision-making.	Technique	(Rooba et al., 2023)

Literature Review

Literature Review	Topic	Citation
Understanding the author's opinion and the user experience is valuable. Opinion mining involves extracting and categorizing opinions from online forums or platforms. This allows for a deeper understanding of user sentiment towards a specific topic.	Technique	(Pavitha et al., 2023)
The linguistic barrier hinders sentimental analysis. Only English reviews can be analyzed. Sarcastic or ironic reviews lead to incorrect classification.	Techniques (Limitation)	(Raut et al., 2023)
Research on sentiment analysis of Indian regional language texts is impeded by the absence of regional language datasets.	Dataset (Limitation)	(Mohan et al., 2023)

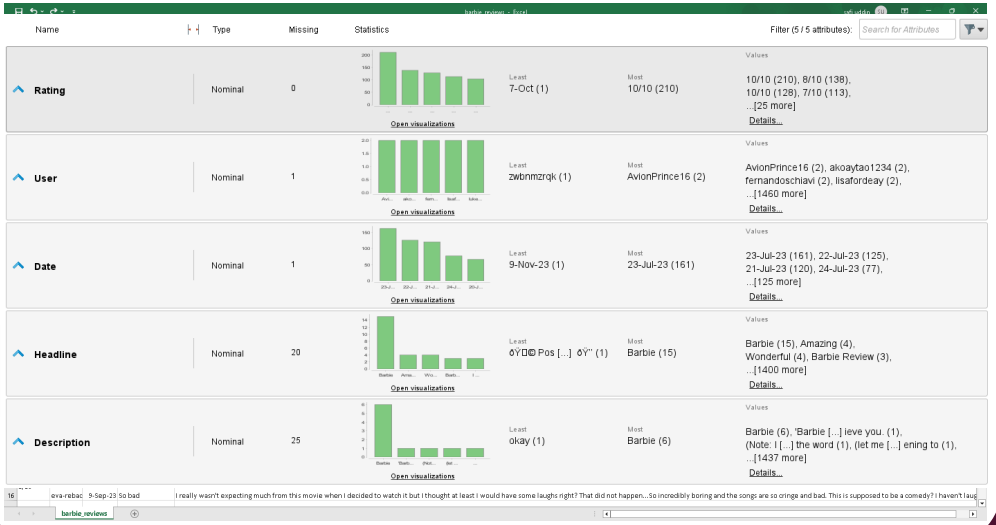
02

Data Understanding

Data collection, and web-scraping

Data Collection

Source of Data: IMDb



Dataset:

Link: https://www.imdb.com/title/tt1517268/reviews

No. of reviews: 1,471

Attributes:

- Rating
- User
- Date
- Headline
- Description

Has **special characters.**

Has **missing values.**

Web-Scraping

```python
from urllib.request import urlopen as uReq
from bs4 import BeautifulSoup as soup
import csv
import requests

def get_reviews(url):
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
    }
    reviews = []

    while True:
        # Make a request to the current URL
        response = requests.get(url, headers=headers)
        page_html = response.text
        page_soup = soup(page_html, "html.parser")

        for item in page_soup.find_all("div", class_="lister-item-content"):
            rating = item.find("span", class_="rating-other-user-rating").text
            user = item.find("span", class_="display-name-link").text
            date = item.find("span", class_="review-date").text
            title = item.find("a", class_="title").text.strip()
            description = item.find("div", class_="text show-more__control").text.strip()

            reviews.append([rating, user, date, title, description])

        # Check if there's a "Load More" button
        load_more_button = page_soup.find("div", class_="load-more-data")
        if load_more_button:
            # Get the "data-key" attribute from the button
            data_key = load_more_button['data-key']

            # Update the URL to include the data-key for the next request
            url = f'https://www.imdb.com/title/tt1517268/reviews/_ajax?sort=curated&dir=desc
        else:
            break  # No more "Load More" button

    return reviews
```

```python
my_url = 'https://www.imdb.com/title/tt1517268/reviews?sort=curated&dir=desc&ratingFilter=5'
reviews = get_reviews(my_url)
# Open a CSV file for writing
csv_filename = 'barbie_reviews5.csv'
with open(csv_filename, 'w', newline='', encoding='utf-8') as csv_file:
    # Create a CSV writer object
    csv_writer = csv.writer(csv_file)

    # Write headers to the CSV file
    csv_writer.writerow(['Rating', 'User', 'Date', 'Headline', 'Description'])

    # Write the data to the CSV file
    csv_writer.writerows(reviews)

print(f"\n++++++++++++++++++++++++++++++++++++++++++++++")
print(f"\nScraped data rating 5/10 has been saved to {csv_filename}")
#check csv file to see the scraped data
```

```
++++++++++++++++++++++++++++++++++++++++++++++

Scraped data rating 5/10 has been saved to barbie_reviews5.csv
```

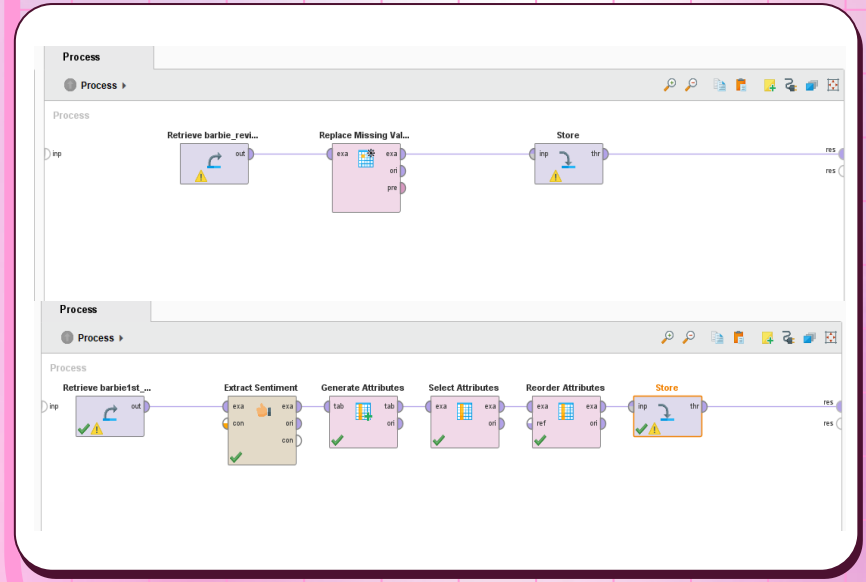
Python

BeautifulSoup

03 Data Preparation

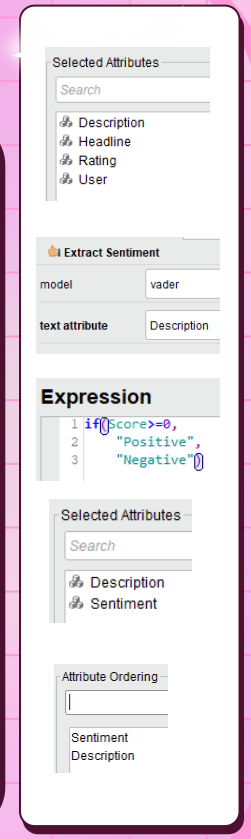
Data Cleaning and Sentiment Extraction

Data Preparation: Cleaning and Extract Sentiment



The Operator

- ***Replace Missing Value***: Remove missing value
- ***Extract Sentiment***: Using Vader technique to analyze the Description
- ***Generate Attributes***: Use IF-Else to create "Sentiment" column for "Positive" and "Negative" attributes.
- ***Select Attributes***: Select only the important attributes. (Sentiment and Description)
- ***Reorder Attributes***: Reorder the attributes position
- ***Store***: Store the data into the repository.

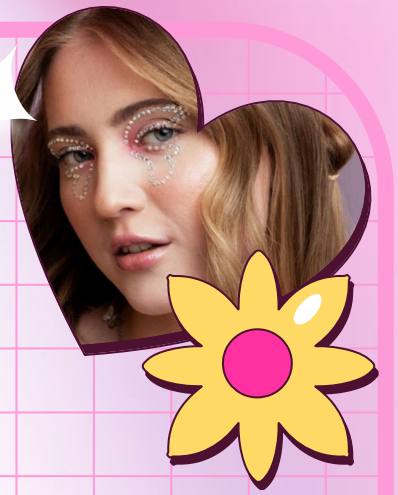


Before

Row No.	Rating	User	Headline	Description
1	1/10	BA_Harrison	Not for me.	Barbie (Marg...
2	1/10	kobalplatt	Did we all se...	I was literally ...
3	1/10	niallmcenteg...	Marketing co...	A word of advi...
4	1/10	STARONTHE...	Overrated	This movie is...
5	1/10	thedarkknight...	Atrocious	Don't know w...
6	1/10	yurixander10	Bad movie.	I bought it on ...
7	1/10	protsenkodv	Barbie Movie ...	Oh boy, wher...
8	1/10	evisadh	Rubbish	Watched it at ...
9	1/10	d-s-allen82	Terrible	This movie w...
10	1/10	ericvankeimp...	Worst movie ...	We saw the ...
11	1/10	owenjmiddlet...	This years fil...	When I went t...
12	1/10	chriswalls-08...	Don't believe ...	Rarely have I ...
13	1/10	mariekeherwi...	Very weak.	I would not re...
14	1/10	pensacolaco...	Really really ...	I am at a com...
15	1/10	eva-rebac	So bad	I really wasn't...
16	1/10	li0904426	Total Disaste...	The movie "B...
17	1/10	bdewiyah	BAD	I am just ama...
18	1/10	amandakelly-...	Rubbish mov...	The only goo...
19	1/10	hamedkhaza...	Failed to be f...	It was one of t...
20	1/10	mr_asdfg	Terribble	The Barbie m...

After

Row No.	Sentiment	Description
1	Positive	Barbie (Marg...
2	Negative	I was literally ...
3	Negative	A word of advi...
4	Negative	This movie is...
5	Positive	Don't know w...
6	Positive	I bought it on ...
7	Positive	Oh boy, wher...
8	Negative	Watched it at ...
9	Positive	This movie w...
10	Positive	We saw the ...
11	Positive	When I went t...
12	Negative	Rarely have I ...
13	Positive	I would not re...
14	Negative	I am at a com...
15	Positive	I really wasn't...
16	Positive	The movie "B...
17	Positive	I am just ama...
18	Positive	The only goo...
19	Positive	It was one of t...
20	Negative	The Barbie m...

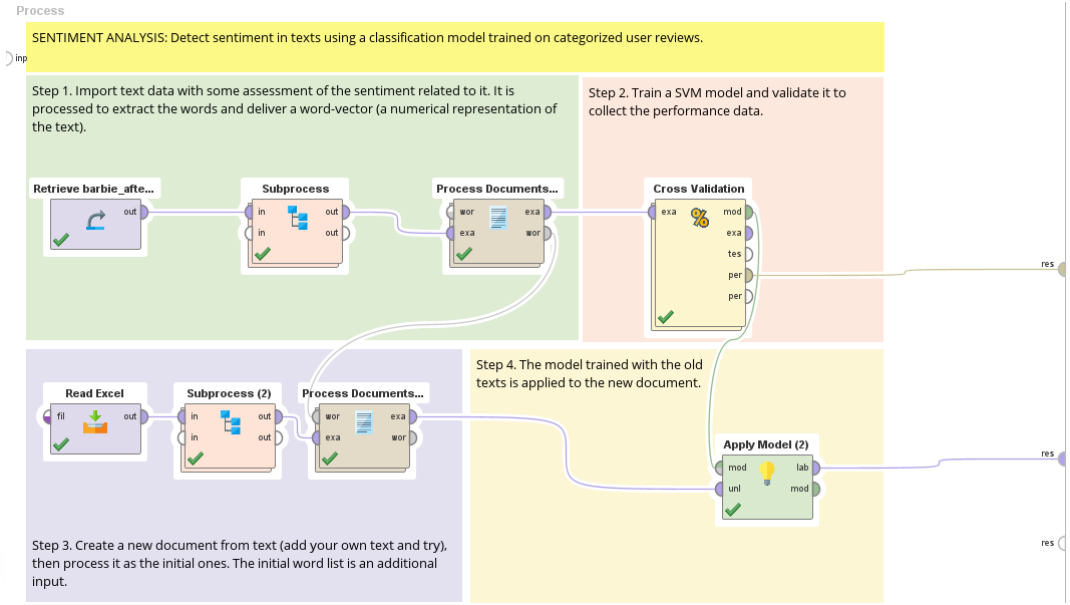


04 Modelling

Naive Bayes, SVM, Random Forest, and Decision Tree

Data modelling

Main Process Structure

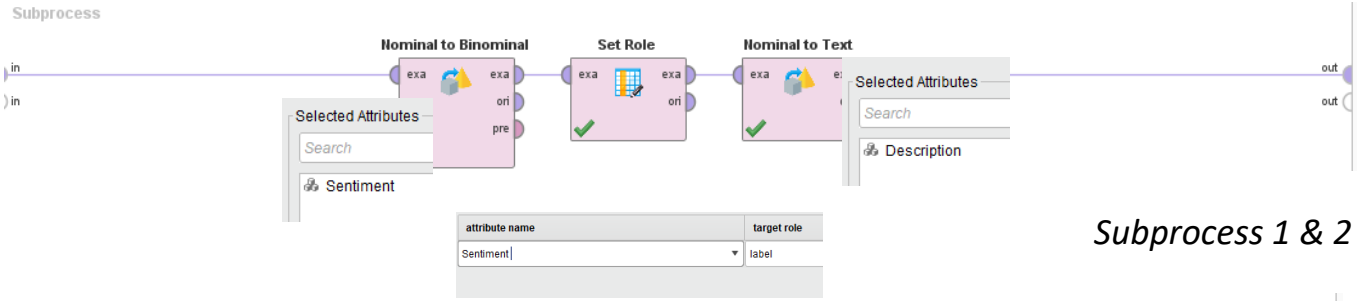


Model

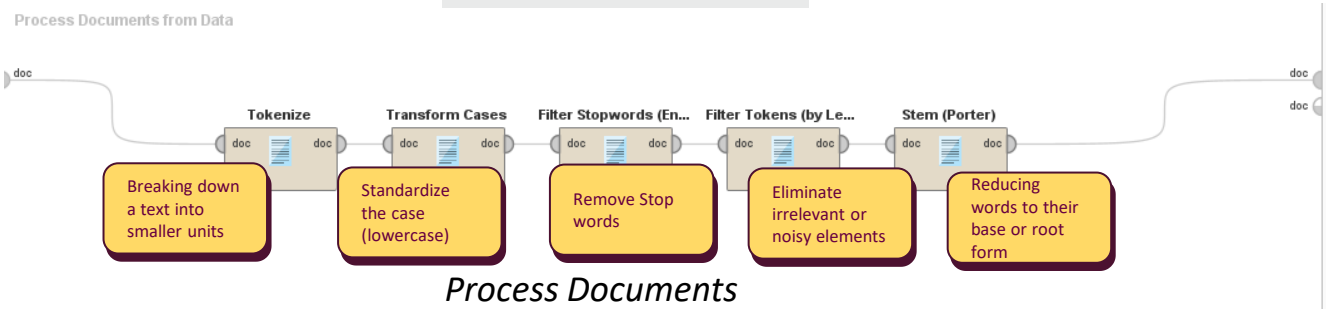
- ***SVM***
- ***Naïve Bayes***
- ***Random Forest***
- ***Decision Tree***

Data modelling

Sub Process Structure



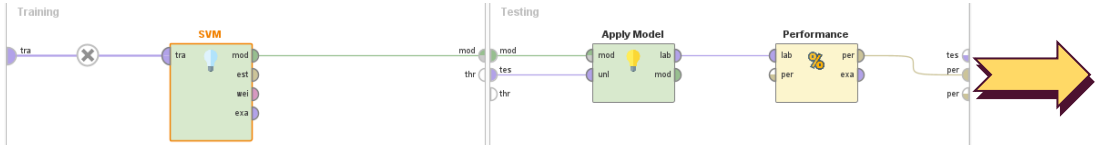
Subprocess 1 & 2

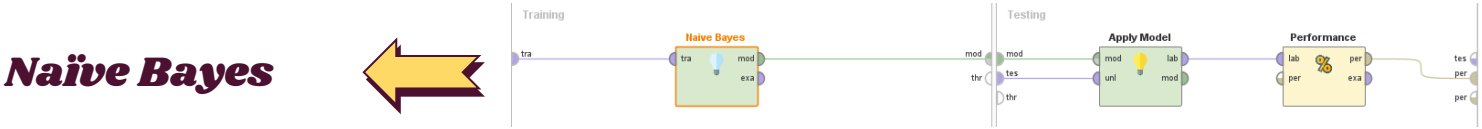


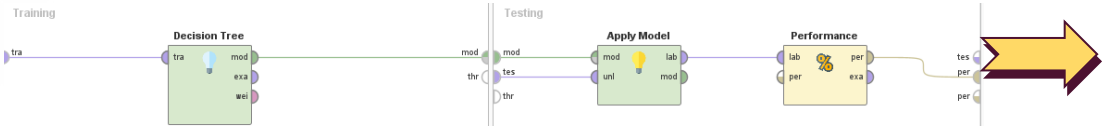
Process Documents

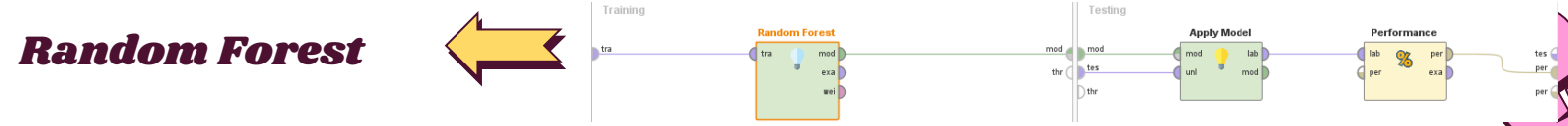
Data modelling

Cross Validation Structure







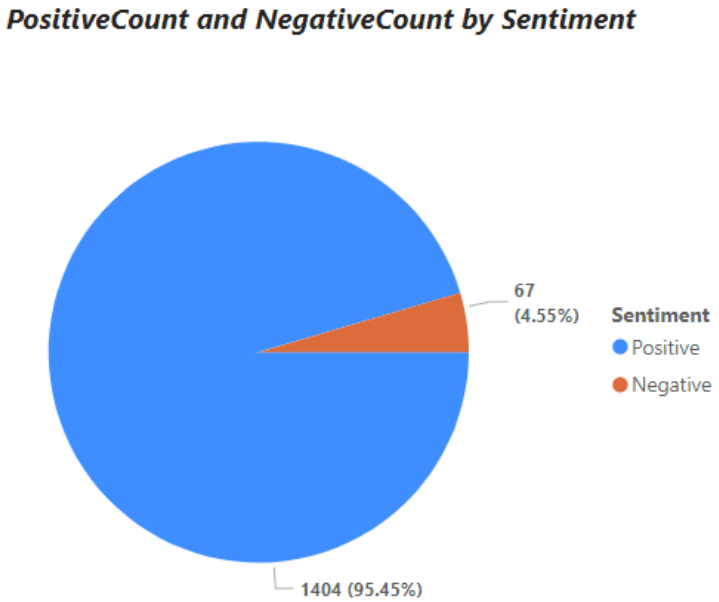


05

Evaluation

Visualize model performance, comparison and sentiment

Results


PositiveCount and NegativeCount by Sentiment
67
(4.55%)
1404 (95.45%)
Sentiment
Positive
Negative

Model

SVM

accuracy: 95.45% +/- 0.33% (micro average: 95.45%)

	true Positive	true Negative	class precision
pred. Positive	1404	67	95.45%
pred. Negative	0	0	0.00%
class recall	100.00%	0.00%	

Naïve Bayes

accuracy: 95.17% +/- 0.60% (micro average: 95.17%)

	true Positive	true Negative	class precision
pred. Positive	1400	67	95.43%
pred. Negative	4	0	0.00%
class recall	99.72%	0.00%	

Model

Random Forest

accuracy: 95.45% +/- 0.33% (micro average: 95.45%)

	true Positive	true Negative	class precision
pred. Positive	1404	67	95.45%
pred. Negative	0	0	0.00%
class recall	100.00%	0.00%	

Decision Tree

accuracy: 94.36% +/- 0.91% (micro average: 94.36%)

	true Positive	true Negative	class precision
pred. Positive	1383	62	95.71%
pred. Negative	21	5	19.23%
class recall	98.50%	7.46%	

Results


PERFOMANCE COMPARISON
Accuracy
Recall
Precision
95.17%
99.72%
94.43%
95.45%
100%
95.45%
95.45%
100%
95.45%
94.36%
98.50%
95.71%
NAÏVE BAYES
SUPPORT VECTOR MACHINE
RANDOM FOREST
DECISION TREE

06

Conclusion

Summarize: Findings and implications

Conclusion

- Achieved all of the objectives
- In this website, majority of the watchers give positive feedback.
- 4 model have been used to analyze the sentiment.
- All model has accuracy, recall, and precision above 90%.
- Random Forest and SVM shows highest accuracy which is 95.45%

References

N Pavithaa, Vithika Pungliyaa, Ankur Raut, Roshita Bhonsle, Atharva Purohit, Aayushi Patel, R Shashidhar. (2023). *Movie recommendation and sentiment analysis using machine learning.* https://doi.org/10.1016/j.gltp.2022.03.012

N. Saraswathi, T. Sasi Rooba, S. Chakaravarthi. (2023). *Improving the accuracy of sentiment analysis using a linguistic rule-based feature selection method in tourism reviews.* https://doi.org/10.1016/j.measen.2023.100888

Juan Kong, Chen Lou. (2023). *Do cultural orientations moderate the effect of online review features on review helpfulness? A case study of online movie reviews.* https://doi.org/10.1016/j.jretconser.2023.103374

Syam Mohan E , R. Sunitha. (2023). *MABSA: A curated Malayalam aspect based sentiment analysis dataset on movie reviews.* https://doi.org/10.1016/j.dib.2023.109452

Thanks!

Do you have any questions?