Data Science Engineering Methods and Tools

Raj Sarode
002762015

Problem Statement & Approach

• Problem Statement: Scrape about 10,000 reviews on any product of your choice and label each review as negative if its rating is 1 or 2, else positive if its rating is 4 or 5. Build an NBC classifier and evaluate the performance with 80/20 split. Please make sure to apply tokenization and stemming for preparing the textual review

• Approach:

- Perform web scrapping using beautiful soup.
- Trained the scraped reviews NBC model on the training set.

• Outcome:

- Perfectly received reviews for a movie.
- Achieved accuracy of after successfully executing the reviews on NBC.

Web Scrapping

	A	
1	Review	Label
2	There is no doubt that the movie was well thought of - from the plot to execution. I must praise Phoenix's superb acting. He managed to	positive
3	The movie affects you in a way that makes it physically painful to experience, but in a good way.	positive
4	When I heard everyone saying that this is the film of the year and all the reviews flooding in with 10's, I was quite hyped and excited for this mo	positive
5	Truly a masterpiece, The Best Hollywood film of 2019, one of the Best films of the decade... And truly the Best film to bring a comic book so	positive
6	I have seen Joker yesterday at Venice an early ill-fated screening. We had some trouble with audio that lead to a near-hour delay, but it	positive
7	I get why some people hate this . It's because of the political message and how some people think that you need get empathy for Arthur's mad	positive
8	Let me start off by saying if Joaquin Phoneix doesn't get his Oscar for this movie. Then the Oscars should be cancelled. Phoneix is amazing as yo	positive
9	Every once in a while a movie comes, that truly makes an impact. Joaquin's performance and scenography in all it's brilliance. Grotesque, haunt	positive
10	This is a movie that only those who have felt alone and isolated can truly relate to it. You understand the motive and you feel sorry for the chara	positive
11	Here, we have a broken man whose mental instability and illness is the catalyst for what Arthur Fleck eventually becomes, but I mean nowhere	negative
12	It's sad that Joaquin missed Oscar for 'The gladiator' as he was very compelling Villain. But I am quite confident he will win it for the 'Joker'.	positive
13	I was so hyped up to see this movie, loved the trailer, only movie I have wanted to watch in the cinema for a very long time.I watched it and wa	positive
14	They say it's all in the writing. And this is a great example of that. Good acting and production values cannot save a script steeped in tiresome st	negative
15	Joaquin Phoenix gives a tour de force performance, fearless and stunning in its emotional depth and physicality. It's impossible to talk about this	positive
16	Need I say more? Everything about this Movie is Masterful in every single way! Joker isn't just an awesome comic book movie, it's an awesome	positive
17	While I've watched many of the superhero movies that have come down the pike,	positive
18	Most of the time movies are anticipated like this they end up falling short, way short. Joker is the first time I was more than happy with the hype	positive
19	I have just watched the Joker in Venice and I will say if Joaquin doesn't get an Oscar this year then something is wrong with this world. This perf	positive
20	In an era of cinema so saturated with superheroes and gritty remakes, Todd Phillips and Phoenix somehow take a character so deeply intrenche	positive
21	I thought this film was good but I just don't get the hype personally. The acting was amazing and the film was good overall but I think 'masterpie	positive
22	The acting, cinematography, sound design, and the script itself is phenomenal. This movie is a triumph. Joaquin Pheonix deserves an Oscar win f	positive
23	The reviews here are making this film out to be far better than it actually is. Don't get me wrong, technically, it is extremely well done and Phoen	positive
24	I know this is an unpopular opinion, but honestly this movie is overrated and boring. 2 hours felt like 3 and a half.I don't buy the whole thing	negative
25		

```
# Loop through the reviews and extract the relevant information such as the review text,
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
corpus = []
labels = []

for review in reviews:
    review_text_elem = review.find('div', {'class': 'text'})
    rating_elem = review.find('span', {'class': 'rating-other-user-rating'})
    date_elem = review.find('span', {'class': 'review-date'})

    # Check if the elements are None before accessing their attributes
    review_text = review_text_elem.text.strip() if review_text_elem is not None else ''
    rating = rating_elem.find('span').text.strip() if rating_elem is not None else ''
    date = date_elem.text.strip() if date_elem is not None else ''

    if rating != '':
        if int(rating) >= 7:
            labels.append('positive')
        else:
            labels.append('negative')
        corpus.append(review_text)
```

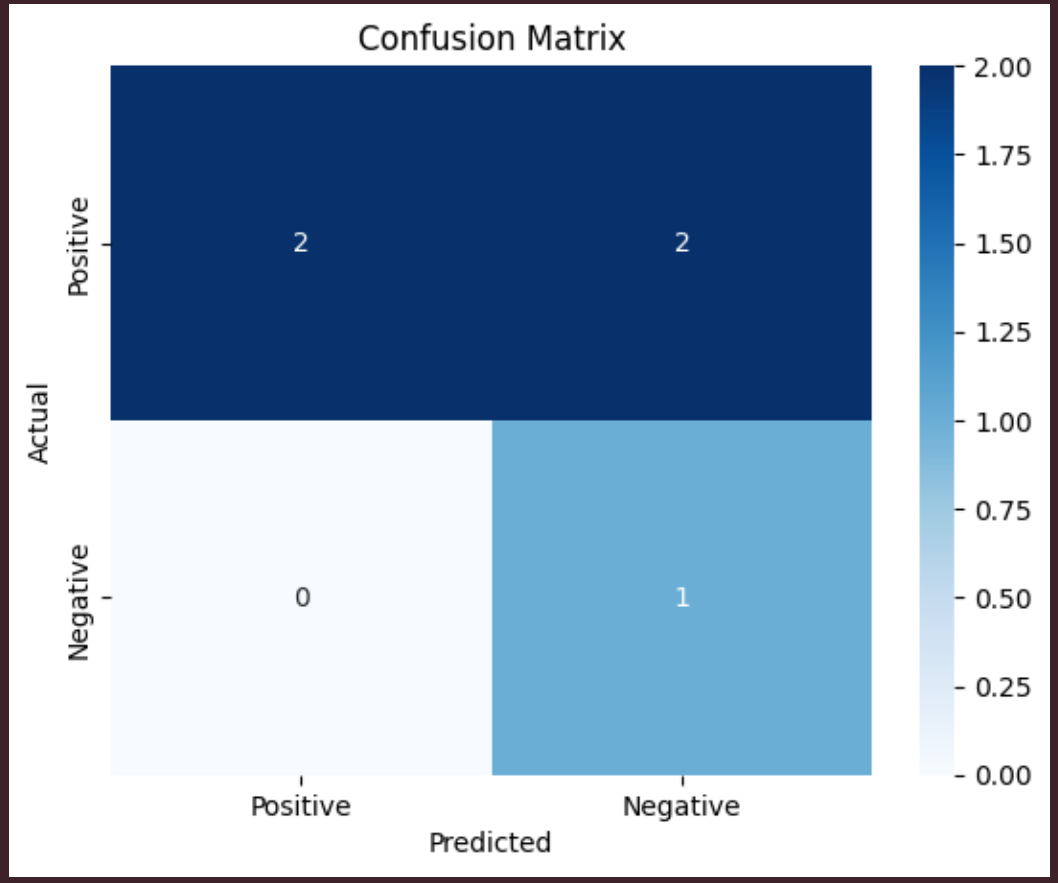
NBC Model

```
# Preprocess the text data using CountVectorizer
vectorizer = CountVectorizer(stop_words='english')
X_train = vectorizer.fit_transform(corpus_train)
X_test = vectorizer.transform(corpus_test)

# Train a Naive Bayes classifier
clf = MultinomialNB()
clf.fit(X_train, labels_train)

# Evaluate the classifier on the testing set
pred_labels_test = clf.predict(X_test)

# Calculate and print the classifier's performance metrics
accuracy = accuracy_score(labels_test, pred_labels_test)
precision = precision_score(labels_test, pred_labels_test, pos_label='positive')
recall = recall_score(labels_test, pred_labels_test, pos_label='positive')
f1 = f1_score(labels_test, pred_labels_test, pos_label='positive')

print('Accuracy:', accuracy)
print('Precision:', precision)
print('Recall:', recall)
print('F1 Score:', f1)
```

Corresponding Accuracies:

```
Accuracy: 0.8
Precision: 0.8
Recall: 1.0
F1 Score: 0.888888888888889
```

Results



Accuracy – 0.6000
Precision – 1.0
Recall – 0.500
F1 Score – 0.6667